12-19-2003

Marginal Modeling of Multilevel Binary Data with Time-Varying Covariates

Diana Miglioretti
Group Health Cooperative, miglioretti.d@ghc.org

Patrick Heagerty
University of Washington, heagerty@u.washington.edu

Abstract. We propose and compare two approaches for regression analysis of multilevel binary data when clusters are not necessarily nested: a GEE method that relies on a working independence assumption coupled with a three-step method for obtaining empirical standard errors, and a likelihood-based method implemented using Bayesian computational techniques. Implications of time-varying endogenous covariates are addressed. The methods are illustrated using data from the Breast Cancer Surveillance Consortium to estimate mammography accuracy from a repeatedly screened population.

KEY WORDS: longitudinal data, endogeneity, conditional, marginal, transition models, hierarchical models.

1 Introduction

Large biomedical data sets often confront investigators with the need to address multiple levels of "clustering" that arise from the organizational structure of the health care delivery system. For example, multiple patients may be evaluated or treated by the same physician. Furthermore, multiple physicians may practice within a clinic or hospital unit and share common beliefs or policies. When large data sets encompass outcomes on individual patients and analysis focuses on the relationship between patient outcomes and measured characteristics of patients, doctors, or clinics, a proper statistical analysis must consider the potential correlation induced by unmeasured heterogeneity at each level of the organizational hierarchy.

A branch of statistics commonly referred to as "multilevel models" (Goldstein, 1995) or "hierarchical linear models" (Bryk and Raudenbush, 1992) has developed in response

to the organizational clustering found in educational settings where students are nested within classrooms, and classrooms are nested within schools. This data structure motivated development of statistical methods that explicitly parameterize systematic components of variation attributable to measured characteristics of both subjects and clusters (*i.e.*, covariates for students and classrooms) and that characterize the magnitude of random or unmeasured heterogeneity as represented by random effects. Although hierarchical models in the educational evaluation literature focused on continuous outcomes and based inference on multivariate normal models, recent interest has considered the extension to discrete outcomes using mixed-effects generalized linear models (Hedeker and Gibbons, 1994; Daniels and Gatsonis, 1999; Rodriguez and Goldman, 2001).

Longitudinal data can also be viewed as a type of multilevel data where repeatedly measured outcomes are clustered within a subject (Diggle, Heagerty, Liang and Zeger, 2002). However, methods specifically developed for the analysis of longitudinal data also explicitly acknowledge the time ordering of measurements and adopt correlation models that capture short-term serial correlation not explained by cluster-level random effects. For example, Diggle (1988) discusses the use of a model with random intercepts and a continuous time auto-regressive error process.

Despite the richness of models and estimation algorithms for continuous outcomes, modeling of multilevel binary data remains a significant challenge in many biomedical applications. Short categorical time series are typical in longitudinal epidemiological studies. Hierarchical models using the standard assumption of normally distributed subject-specific effects can be difficult to fit and may not adequately characterize the multivariate categorical structure (see Carlin *et al.*, 2001; Agresti and Liu, 1999). When substantive interest is in the

marginal regression structure, conditionally specified generalized linear mixed models do not directly address the scientific question, and must either be marginalized to obtain model summaries or reparameterized to allow direct inference on marginal contrasts. See Chapter 7 of Diggle, Heagerty, Liang, and Zeger (2002) for a comparison and discussion of marginal and conditional approaches.

A generalized estimating equations (GEE; Liang and Zeger, 1986) approach directly models the marginal mean and may be computationally feasible even with large numbers of subjects. However, without modification this approach may give biased results when data are not missing completely at random (Laird, 1988; Robins *et al.*, 1995). Estimation that relies on a working independence correlation structure may be less efficient than a correctly specified maximum likelihood estimator, because efficient inference using GEE demands that the working correlation model approximates the true correlation structure of the data (Wang and Carey, 2003; Shults and Morrow, 2002). In addition, directly using GEE for non-nested clusters or incompletely crossed designs has not been previously investigated. Betensky *et al.* (2000) propose "reclustering" by grouping observations into independent blocks of data, but this is not feasible for crossed designs. Reclustering may also lead to a small number of independent blocks which is known to produce negatively biased sandwich variance estimates (Mancl and DeRouen, 2001).

Further complications arise in the case of stochastic time-varying covariates. A time-varying covariate is exogenous when it is not predicted by past outcomes. Formally, under an exogenous covariate process, $p\left(x_t|H_{t-1}\left(y\right),H_{t-1}\left(x\right)\right)=p\left(x_t|H_{t-1}(x)\right)$ where $H_t\left(u\right)$ is the history of u up to and including time t. In contrast, an endogenous covariate is conditionally dependent on past response values. See Diggle, Heagerty, Liang, and Zeger (2002) Chapter 12

for more detailed discussion. Hierarchical models and GEE with weighted working covariance matrices assume that the full covariate conditional mean, i.e., the mean given the covariate vector from all time points, is equal to the cross-sectional mean (Pepe and Anderson, 1994; Diggle, Heagerty, Liang and Zeger, 2002). This assumption can be met for an exogenous covariate process by including appropriate current or lagged values of the covariate. When covariates are endogenous, GEE with a working independence covariance matrix may be used to characterize the cross-sectional or lagged association, but the necessary use of diagonal covariance weighting may result in a sacrifice of efficiency. With endogenous treatment or exposure variables, alternative causal inference targets and methods of estimation have been proposed. Robins, Greenland, and Hu (1999) discuss targets of inference and contrast their causal estimation methods with standard regression approaches that simply characterize associations among observed random variables. In our motivating example we are interested in descriptive models for assessing systematic variation in the accuracy of screening tests rather than making causal inference statements based on potential outcomes.

Given the complementary advantages and assumptions of GEE and likelihood-based methods, we propose two marginal approaches to account for the correlation within short time series measured on individual subjects, and the correlation induced through organizational clustering of patients within a doctor. The first approach is a marginalized multilevel model based on the ideas described in Heagerty and Zeger (2000), Heagerty (2002), and Diggle, Heagerty, Liang, and Zeger (2002). This approach combines a marginal generalized linear model that estimates the influence of covariates on the marginal probability of a positive response with a conditional logistic regression model that describes the dependence structure. The conditional model captures the serial dependence within individuals

using a Markov structure and includes cluster-specific effects to account for the correlation within the larger clusters. The second approach is a three-step strategy for fitting GEE to non-nested clusters using standard software.

This work was motivated by a large multi-site study aiming to estimate the accuracy of screening mammography as practiced in the community and to describe how the accuracy varies across different subgroups of women. Women are screened at multiple time points, and outcomes are correlated within radiologists, who typically screen hundreds to thousands of women annually. Women are not necessarily nested within radiologists. The accuracy of mammography is characterized by its sensitivity, the prevalence of a positive/abnormal mammogram result among woman with breast cancer, and its specificity, the prevalence of a negative/normal mammogram result among woman without breast cancer. Marginal logistic regression models provide a convenient and direct approach for modeling changes in sensitivity and specificity across sub-populations defined by measured covariate values.

In the next section, we introduce the marginalized multilevel model and describe model fitting using a Bayesian approach under the assumption of exogenous covariates and in the case of endogenous covariates. In section 3, we describe a three-step estimation strategy based on GEE as implemented in standard software. We illustrate the approach using data from the Breast Cancer Surveillance Consortium in Section 4. We conclude with a comparison of the methodologies.

2 Marginalized Multilevel Model

We consider the case where interest is in the marginal effects of covariates on the probability of a repeatedly measured binary outcome that is clustered, but not necessarily nested, within an additional level (*e.g.*, repeatedly screened individuals clustered within radiologists). Extensions to more than two clustering levels is straightforward. Let y_{it} be the tth binary outcome for the ith individual; $i = 1, \ldots, N; t = 1, \ldots, T_i$. We model the influence of a $p \times 1$ vector of possibly time-varying covariates $\mathbf{x}_{it}$ on the marginal probability of a positive response μ_{it}^M using logistic regression:

$$\begin{aligned} \mu_{it}^M &= p\left(y_{it} = 1 | H_{it}(\mathbf{x})\right) \\ \text{logit}\left(\mu_{it}^M\right) &= \mathbf{x}_{it}\boldsymbol{\beta} \end{aligned} \tag{1}$$

where $H_{it}(\mathbf{x}) = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it})$ denotes covariates measured up to and including time t. Here, we assume the regression model properly specifies the conditional mean given the covariate history (Pepe and Andersen, 1994) such that $E\left(y_{it}|\mathbf{x}_{it}\right) = E\left(y_{it}|H_{it}(\mathbf{x})\right)$. In general, this condition assumes that stochastic time-varying covariates are properly modelled through $\mathbf{x}_{it}$, possibly by including lagged covariates.

Let c_{it} indicate the cluster to which the ith individual belongs at time t. We do not assume individuals are nested within a cluster over time. For example, a woman's mammograms may be read by different radiologists at different visits. We capture the dependence structure of $\mathbf{y}_i$ through a conditional logistic regression model that includes the previous outcome y_{it-1} to account for the serial correlation within individuals and incorporates cluster-specific effects

u_j to account for correlation within larger clusters $c_{it}; j = 1, \ldots, J$:

$$\begin{aligned} \mu_{it}^C &= p\left(y_{it} = 1 | \mathbf{x}_{it}, y_{it-1}, c_{it} = j, u_j\right) \\ \text{logit}\left(\mu_{it}^C\right) &= \Delta_{it} + \gamma_t y_{it-1} + u_j \\ u_j &\sim N\left(0, 1/\tau\right) \end{aligned} \tag{2}$$

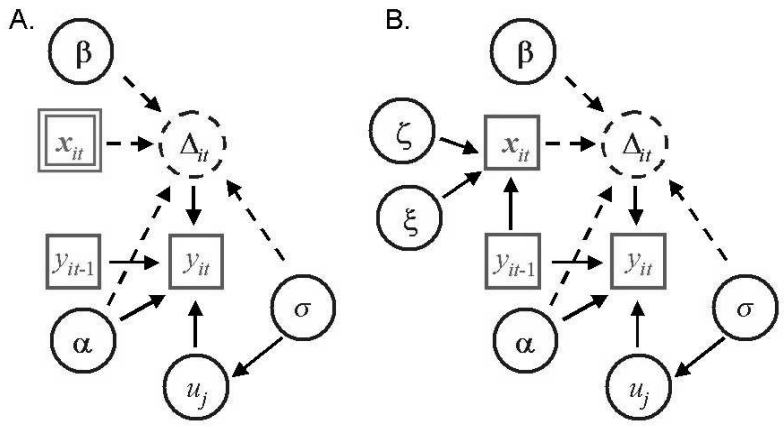
The individual-specific intercept Δ_{it} is fully constrained by the relationship between the marginal and conditional means, as described below. We assume a first-order Markov model since we are dealing with short time series; however, higher order models may be adopted (Heagerty, 2002). A regression model may also be specified for γ_t :

$$\gamma_t = \mathbf{z}_t \boldsymbol{\alpha}$$

where $\boldsymbol{\alpha}$ measures how the dependence of y_{it} on y_{it-1} varies as a function of covariates $\mathbf{z}_t$.We take the cluster-specific effects u_j to be normally distributed since u_j represents the average effect of all unmeasured/unobservable cluster-specific factors. By the central limit, this average additive effect will tend toward normality as the number of latent covariate effects increases.

The marginalized multilevel model has several advantages. First, the marginal mean is directly modeled so the regression coefficients β have cross-sectional or population average interpretations. Second, the mean model is separate from the association model. As a result, the interpretation of the regression coefficients β does not depend on the specification of the association model. Last, the dependence within women is modeled using a transition

model, which is a natural specification for short, binary time series. For further discussion on marginalized and conditional multilevel models, see Heagerty and Zeger (2000).



Directed acyclic graphical models for y_{it} when covariates $\mathbf{x}_{it}$ are exogenous (A) and endogenous (B).

Figure 1 shows the directed acyclic graphical models (DAG) for y_{it} in the cases of fixed (and thus exogenous) covariates (Figure 1A) and endogenous covariates (Figure 1B). Note that stochastic time-varying covariates may also be exogenous, but we have represented the covariates in Figure 1A as fixed for simplicity. Unknown parameters are represented by solid circles. Dashed circles represent deterministic functions of these parameters. Single squares represent observed random variables and double squares represent fixed covariates. Solid arrows, drawn from parent nodes to their descendents, represent probabilistic dependences and dashed arrows show deterministic relationships. We assume $\mathbf{x}$ is independent of $\mathbf{u}$ since individual-level covariates should not depend on radiologist-specific effects. The key distinction between Figures 1A and 1B is that under an endogenous covariate process, the previous response y_{it-1} predicts future covariates $\mathbf{x}_{it}$. The parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ describe the covariate submodel, which is discussed in subsection 2.2.

2.1 Likelihood-based Estimation with Exogenous Covariates

Assuming outcomes within an individual are independent conditional on the previous result y_{it-1} and cluster-specific effects $\mathbf{u}$, the likelihood may be written as follows:

$$p\left(\mathbf{y}|\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{u},\tau\right) \propto \prod_{i=1}^{N}\prod_{t=1}^{T_i}\left(\mu_{it}^{C}\right)^{y_{it}}\left(1-\mu_{it}^{C}\right)^{(1-y_{it})}$$

where $\mu_{it}^{C} = \text{logit}^{-1}\left(\Delta_{it} + \gamma_t y_{it-1} + u_j\right)$.

We use Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution, which is proportional to the product of the prior distributions and the likelihood:

$$p\left(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{u},\tau|\mathbf{y},\mathbf{x}\right) \propto p\left(\boldsymbol{\alpha}\right)p\left(\boldsymbol{\beta}\right)p\left(\tau\right)\prod_{j=1}^{R}p\left(u_j|\tau\right)\prod_{i=1}^{N}\prod_{t=1}^{T_i}\left(\mu_{it}^{C}\right)^{y_{it}}\left(1-\mu_{it}^{C}\right)^{(1-y_{it})}.$$

We present our fitting approach using standard prior distributions, taking $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be $normal\left(0, 1/\psi\right)$ and τ to be $gamma\left(A, B\right)$ with investigator-specified hyperparameters ψ, A, and B. In the style of Gibbs sampling, each set of parameters is updated conditional on the remaining parameters using Metropolis (random walk) steps (Metropolis, *et al.*, 1953; Gilks, Richardson, and Spiegelhalter, 1996). The algorithm along with acceptance probabilities are detailed in Appendix 1.

To calculate the likelihood, we need to determine the values of $\boldsymbol{\Delta}$ such that equations

(1) and (2) are simultaneously satisfied. To do this, we use the following relationship:

$$\begin{aligned}\mu_{it}^{M} &= E_{\mathbf{u}}\left\{E_{y_{it-1}}\left[E\left(y_{it}|y_{it-1},H_{it}(\mathbf{x}),\mathbf{u}\right)\right]\right\} \qquad (3)\\ &= \begin{cases} \int\left\{h\left(\Delta_{it},1,z\right)\mu_{it-1}\left(z\right)+h\left(\Delta_{it},0,z\right)\left[1-\mu_{it-1}\left(z\right)\right]\right\}\phi\left(z\right)dz & \text{if } c_{it}=c_{it-1}\\ \int\left\{h\left(\Delta_{it},1,z\right)\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)+h\left(\Delta_{it},0,z\right)\left[1-\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)\right]\right\}\phi\left(z\right)dz & \text{if } c_{it}\neq c_{it-1}\end{cases}\end{aligned}$$

where $h\left(\Delta_{it},y_{it-1},z\right) = \text{logit}^{-1}\left(\Delta_{it}+\gamma_t y_{it-1}+\sigma z\right)$, $\mu_{it-1}\left(z\right) = p\left(y_{it-1}=1|\mathbf{x}_{it},H_{it-1}(\mathbf{x}),u_j=\sigma z\right)$, $\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right) = p\left(y_{it-1}=1|\mathbf{x}_{it},H_{it-1}(\mathbf{x})\right)$, $\sigma = \sqrt{1/\tau}$, and $\phi\left(z\right)$ is the standard normal density.

To solve for Δ_{it} when $t>1$ we need the values of $\mu_{it-1}\left(z\right)$ and $\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)$. Under the assumption of exogenous covariates $\mathbf{x}_{it}$; *i.e.*, $p\left(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT_i}\right)=p\left(y_{it}|H_{it}(\mathbf{x})\right)$; $\mu_{it-1}\left(z\right)$ and $\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)$ do not depend on $\mathbf{x}_{it}$ given $H_{it-1}(\mathbf{x})$ (Diggle, Heagerty Liang, and Zeger, 2002). In this case, $\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)$ =$\text{logit}^{-1}\left(\mathbf{x}_{it-1}\boldsymbol{\beta}\right)$ and we can easily calculate $\mu_{it}\left(z\right)$ by first solving for Δ_{i1} and $\mu_{i1}\left(z\right)$ and then sequentially updating Δ_{it} and $\mu_{it}\left(z\right)$ given $\mu_{it-1}\left(z\right)$ and $\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)$. Details are presented in Appendix 2.

2.2 Likelihood-based Estimation with an Endogenous Covariate

In the previous estimation algorithm for an exogenous covariate processes, we could assume $p\left(y_{it-1}|\mathbf{x}_{it},H_{it-1}(\mathbf{x})\right) = p\left(y_{it-1}|H_{it-1}(\mathbf{x})\right)$, and this equality was used to marginalize $\mathbf{y}_{it}$ and solve for Δ_{it}. However, if the covariate process is endogenous, $\mathbf{x}_{it}$ depends on y_{it-1} and therefore $p\left(y_{it-1}|\mathbf{x}_{it},H_{it-1}(\mathbf{x})\right)$ will no longer equal $p\left(y_{it-1}|H_{it-1}(\mathbf{x})\right)$. In order to marginalize in this situation, we model the covariate process in addition to the response process, which allows recovery of Δ_{it} and evaluation of the likelihood.

In the case of an endogenous covariate, $\mu_{it}(z)$ and $\mu^M_{it-1}(H_{it}(\mathbf{x}))$ can be estimated through the following factorizations:

$$\begin{aligned}\mu_{it-1}(z) &= p(y_{it-1}|\mathbf{x}_{it}, H_{it-1}(\mathbf{x}), u_j) = \frac{p(y_{it-1}|H_{it-1}(\mathbf{x}), u_j)\, p(\mathbf{x}_{it}|y_{it-1}, H_{it-1}(\mathbf{x}))}{p(\mathbf{x}_{it}|H_{it-1}(\mathbf{x}))} \quad (4)\\ \mu^M_{it-1}(H_{it}(\mathbf{x})) &= p(y_{it-1}|\mathbf{x}_{it}, H_{it-1}(\mathbf{x})) = \frac{p(y_{it-1}|H_{it-1}(\mathbf{x}))\, p(\mathbf{x}_{it}|y_{it-1}, H_{it-1}(\mathbf{x}))}{p(\mathbf{x}_{it}|H_{it-1}(\mathbf{x}))}\end{aligned}$$

where

$$\begin{aligned}p(\mathbf{x}_{it}|H_{it-1}(\mathbf{x})) &= p(\mathbf{x}_{it}|y_{it}=1, H_{it-1}(\mathbf{x}))\, p(y_{it-1}=1|H_{it-1}(\mathbf{x}), u_j=\sigma z)\\ &\quad + p(\mathbf{x}_{it}|y_{it}=0, H_{it-1}(\mathbf{x}))\, p(y_{it-1}=0|H_{it-1}(\mathbf{x}), u_j=\sigma z).\end{aligned}$$

and we make the reasonable assumption that $p(\mathbf{x}_{it}|y_{it-1}, H_{it-1}(\mathbf{x}))$ and $p(\mathbf{x}_{it}|H_{it-1}(\mathbf{x}))$ do not depend on $\mathbf{u}$. We may estimate $p(\mathbf{x}_{it}|y_{it-1}, H_{it-1}(\mathbf{x}))$ by fitting a generalized linear model for each endogenous covariate x_{itk}:

$$g[E(x_{itk}|y_{it-1}, H_{it-1}(\mathbf{x}))] = \zeta_{0k} + \zeta_{1k} y_{it-1} + \boldsymbol{\xi}_k H_{it-1}(\mathbf{x}).$$

We now need to sample from the full posterior distribution:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \tau, \boldsymbol{\zeta}, \boldsymbol{\xi}|\mathbf{y}, \mathbf{x}) \propto p(\boldsymbol{\alpha})\, p(\boldsymbol{\beta})\, p(\tau)\, p(\boldsymbol{\zeta})\, p(\boldsymbol{\xi})\, p(\mathbf{y}, \mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \tau, \boldsymbol{\zeta}, \boldsymbol{\xi}).$$

where the joint likelihood for both the response and covariate processes may be factored as

follows:

$$p(\mathbf{y},\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{u},\tau,\boldsymbol{\zeta},\boldsymbol{\xi}) = \prod_{t=1}^{T} p\left[\mathbf{y}_t|\mathbf{x}_t,\boldsymbol{\Delta}_t\left(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\xi},\tau\right),\boldsymbol{\alpha},\mathbf{u}\right] p\left(\mathbf{x}_t|\mathbf{y}_{t-1},H_{t-1}(\mathbf{x}),\boldsymbol{\zeta},\boldsymbol{\xi}\right).$$

Model fitting details may be found in Appendix 2.

3 Generalized Estimating Equations Approach

Previous literature has shown that GEE can be validly applied for estimation of the association between a stochastic time-varying covariate and a longitudinal response if a working independence correlation matrix is used (see Pepe & Anderson 1994 and Diggle, Heagerty, Liang, and Zeger, 2002, section 12.3). Robins et al. (1999) has shown that standard estimation methods such as GEE, while able to validly estimate associations, may not characterize causal effects for time-varying treatments or exposures particularly when the exposures of interest are endogenous. In cancer screening regression is used to structure the cross-sectional association between current disease status and current test result, and therefore GEE with working independence provides a valid analytical method for estimation and inference regarding response and covariate association when time-varying covariates are either exogenous or endogenous. In this section we discuss how standard GEE methods can be exploited to obtain valid inference for regression analysis with non-nested clusters.

To detail the approach we consider use of working independence which solves the estimating equation

$$\sum_i \sum_t D_{it}^T V_{it}^{-1}(y_{it} - \mu_{it}^M) = 0$$

where $D_{it} = \partial \mu_{it}^M / \partial \beta$, and $V_{it} = \text{var}(y_{it} \mid \mathbf{x}_{it})$. Based on results of Mayer-Hamblett and Self (2002) and Lumley and Mayer-Hamblett (2003), the solution to the estimating equations, $\widehat{\beta}$, has an asymptotic variance given as

$$\begin{aligned} \text{var}(\widehat{\beta}) &= A_{N,n}^{-1} B_{N,n} A_{N,n}^{-1} \\ A_{N,n} &= \sum_i \sum_t D_{it}^T V_{it}^{-1} D_{it} \\ B_{N,n} &= \text{var}\left(\sum_i \sum_t U_{it} \right) \end{aligned}$$

where $U_i = D_{it}^T V_{it}^{-1} (y_{it} - \mu_{it}^M)$ and $n = \max_i (T_i)$. A consistent estimate of $B_{N,n}$ can be obtained using

$$\widehat{B}_{N,n} = \sum_i \sum_t \sum_j \sum_s \delta(i,j,s,t) \cdot U_{it} U_{js}^T$$

where $\delta(i,j,s,t) = 1$ if either $i = j$ indicating that observations y_{it} and y_{js} are from the same individual or if $c_{it} = c_{js}$ indicating a common radiologist for the observations, and 0 otherwise.

The indicator $\delta(i,j,s,t)$ can be viewed as a logical "or" operator that captures the *same individual* or the *same radiologist*, and as such can be represented as:

$$\delta(i,j,s,t) = \delta^S(i,j) \; + \; \delta^R(i,j,s,t) \; - \; \delta^S(i,j) \cdot \delta^R(i,j,s,t)$$

where $\delta^S(i,j) = 1$ if $i = j$ and 0 otherwise, and $\delta^R(i,j,s,t) = 1$ if $c_{it} = c_{js}$ and 0 otherwise.

This representation shows that the estimate $\widehat{B}_{N,n}$ can be formed from three contributions:

$$\begin{aligned}
\widehat{B}_{N,n} &= \widehat{B}^{S}_{N,n} + \widehat{B}^{R}_{N,n} - \widehat{B}^{SR}_{N,n} \\
\widehat{B}^{S}_{N,n} &= \sum_i \sum_t \sum_j \sum_s \delta^S(i,j) \cdot U_{it} U^T_{js} = \sum_i \sum_t \sum_s U_{it} U^T_{is} \\
\widehat{B}^{R}_{N,n} &= \sum_i \sum_t \sum_j \sum_s \delta^R(i,j,s,t) \cdot U_{it} U^T_{js} \\
\widehat{B}^{SR}_{N,n} &= \sum_i \sum_t \sum_j \sum_s \delta^S(i,j) \cdot \delta^R(i,j,s,t) \cdot U_{it} U^T_{js} \\
&= \sum_i \sum_t \sum_s \delta^R(i,j,s,t) \cdot U_{it} U^T_{is}
\end{aligned}$$

Operationally this implies that $\widehat{B}_{N,n}$ can be obtained from three standard GEE estimates using: cluster on a variable S-ID that identifies subjects to obtain $\widehat{B}^{S}_{N,n}$; cluster on a variable R-ID that identifies radiologists to obtain $\widehat{B}^{R}_{N,n}$; cluster on a variable SR-ID that identifies unique subject-radiologist combinations to obtain $\widehat{B}^{SR}_{N,n}$.

Using working independence the final estimated variance for $\widehat{\beta}$ is simply a linear combination of variance estimates produced by GEE:

$$\begin{aligned}
\mathrm{var}(\widehat{\beta}) &= A^{-1}_{N,n} B_{N,n} \cdot A^{-1}_{N,n} \\
&= \left(A_{N,n} B^{S}_{N,n} \cdot A^{-1}_{N,n}\right) + \left(A_{N,n} B^{R}_{N,n} \cdot A^{-1}_{N,n}\right) - \left(A_{N,n} B^{SR}_{N,n} \cdot A^{-1}_{N,n}\right) \\
&= V^{S}_{N,n} + V^{R}_{N,n} - V^{SR}_{N,n}
\end{aligned}$$

where $V^{S}_{N,n}$ is the estimated variance from a working independence GEE clustering on S-ID while similarly $V^{R}_{N,n}$ clusters on R-ID and $V^{SR}_{N,n}$ clusters on SR-ID.

To illustrate a non-nested correlation structure and to show why the empirical variance

calculation involves three terms, we present a simple example in Table 1 representing observations obtained from three individuals seen by two radiologists. The correlation between a pair of observations from the same subject is represented by $\times$ and is taken into account when clustering on subject using $\delta^S(i,j)$ to include the cross-product $U_{is}U_{jt}^T$ in the empirical variance calculation given by $\widehat{B}_{N,n}^S$. Correlation between observations on different subjects who see the same radiologist is represented by $\bigcirc$ and is properly accounted for by clustering on radiologist using $\delta^R(i,j,s,t)$ to form the sum $\widehat{B}_{N,n}^R$. Correlations for observations that are from the same subject and the same radiologist are represented by $\otimes$. For example, the first two observations in Table 1 represent measurements obtained at two different time points but for the same subject, and with reading by the same radiologist. The cross-product $U_{is}U_{jt}^T$ for this observation is included in both $\widehat{B}_{N,n}^S$ and $\widehat{B}_{N,n}^R$. By subtracting $\widehat{B}_{N,n}^{SR}$ in the final empirical variance calculation, the "double counting" represented in Table 1 by the symbol $\otimes$ is corrected. The SR-ID for this example is formed by concatenating the S-ID with the R-ID to obtain SR-ID=(11, 11, 21, 22, 22, 31), and identifies four groups of observations.

Table 1. *Example correlation structure for two non-nested clusters. Correlation between observations from the same subject (S-ID) is represented by $\times$. Correlation between observations from the same radiologist (R-ID) is represented by $\bigcirc$. Correlation between observations from the same subject and radiologist are represented by $\otimes$. Blank spaces represent independent observations.*

Observation			1	2	3	4	5	6
	S-ID		1	1	2	2	2	3
		R-ID	1	1	1	2	2	1
1	1	1	$\otimes$	$\otimes$	$\bigcirc$			$\bigcirc$
2	1	1	$\otimes$	$\otimes$	$\bigcirc$			$\bigcirc$
3	2	1	$\bigcirc$	$\bigcirc$	$\otimes$	$\times$	$\times$	$\bigcirc$
4	2	2			$\times$	$\otimes$	$\otimes$	
5	2	2			$\times$	$\otimes$	$\otimes$	
6	3	1	$\bigcirc$	$\bigcirc$	$\bigcirc$			$\otimes$

In this section we have outlined a moment-based approach for estimating regression relationships with time-varying covariates under a non-nested correlation structure. In contrast to the marginalized multilevel model estimated with likelihood-based methods, the GEE approach using working independence does not explicitly parameterize the correlation structure, but rather relies on an empirical variance estimator to non-parametrically capture within-subject and within-reader dependence.

4 Example

We illustrate the proposed approaches using data collected between 1996 and 2000 by a mammography registry that participates in Breast Cancer Surveillance Consortium (BCSC; http://breastscreening.cancer.gov). The BCSC is a NCI-sponsored collaboration between seven population-based mammography registries in the United States, established in 1994 to evaluate the performance of mammography in community settings and to improve our understanding of the effects of screening on cancer outcomes. Each registry prospectively collects demographic, risk-factor, and clinical information each time a woman goes to a participating facility for a mammogram. In addition, each mammography registry links women in their registry to a state tumor registry or regional Surveillance, Epidemiology, and End Results (SEER) program and possibly to pathology databases to collect information on cancer status.

Interest is in estimating the marginal sensitivity and specificity of screening mammography as it is practiced in the community by age, breast density, and whether or not the mammogram was the woman's first mammogram. We considered a mammogram to be

positive if the radiologist gave it a BI-RADS assessment of 0 (needs additional imaging), 4 (suspicious abnormality), 5 (highly suggestive of malignancy), or 3 (probably benign finding) with a recommendation for immediate follow-up. A woman was considered to have breast cancer if she was diagnosed with invasive carcinoma or ductal carcinoma *in-situ* within a year after her mammogram and before her next screening mammogram.

4.1 Model for Mammography Accuracy

Let y_{it} be the mammogram result for the ith woman at her tth screening during follow-up and let d_{it} be her corresponding breast cancer status such that $d_{it} = 1$ if she is diagnosed with breast cancer within the follow-up period and $d_{it} = 0$ if she is cancer free; $i = 1, \ldots, N; t = 1, \ldots, T_i$. Note that t here corresponds to a woman's observation number in the data set, not necessarily the number of mammograms in her lifetime. We jointly model sensitivity and specificity in a single logistic regression model, modeling the marginal probability of a positive mammogram μ_{it}^M as a function of a $p \times 1$ vector of covariates x_{it} and cancer status d_{it} (Pepe, 2003):

$$\text{logit}\left(\mu_{it}^M\right) = \beta_0 + x_{it1}\beta_1 + \cdots + x_{itp}\beta_p + d_{it}\delta_0 + x_{it1}d_{it}\delta_1 + \cdots + x_{itp}d_{it}\delta_p.$$

Sensitivity is defined as the true positive rate or the probability of a positive mammogram given cancer in the follow-up period: $p\left(y_{it} = 1 | d_{it} = 1\right)$. Specificity is one minus the false positive rate or the probability of a negative exam given no cancer in the follow-up period: $1 - p\left(y_{it} = 1 | d_{it} = 0\right)$. The β coefficients capture the influence of covariates on the probability of a positive mammogram (the "call back" rate). The δ coefficients capture the additional

influence of covariates given $d_{it} = 1$. Thus, a test of $\delta = 0$ tests whether x influences the *accuracy* of mammography.

For the marginalized multilevel model, we capture the dependence structure through a conditional logistic regression model that includes the previous outcome y_{it-1} to account for the serial correlation within women and that incorporates radiologist-specific effects. The accuracy of screening mammography is typically only estimated for women without a previous history of breast cancer, since women with a history of breast cancer undergo surveillance mammography as opposed to true screening. Therefore, observation t for woman i is only included in the analysis if $d_{it'} = 0$ for $t' = 1, \ldots, t-1$, and hence, the previous mammogram result is either a true negative (if $y_{it-1} = 0$) or a false positive (if $y_{it-1} = 1$). Let u_j be the effect associated with the jth radiologist, $j = 1, \ldots, J$. To take into account the correlation within women and radiologists, the conditional probability of a positive mammogram μ_{it}^C is modeled as a function of the previous result and the radiologist-specific effect:

$$\begin{aligned} \mu_{it}^C &= p\left(y_{it} = 1 | d_{it}, y_{it-1} \text{ for } t > 1, c_{it} = j, u_j\right) \\ \text{logit}\left(\mu_{it}^C\right) &= \Delta_{it} + \left(\alpha_1 + \alpha_2 d_{it}\right) y_{it-1} + u_j \\ u_j &\sim N\left(0, 1/\tau\right) \end{aligned} \tag{6}$$

An interaction between y_{it-1} and d_{it} was included, since the influence of the previous outcome may depend on current disease status. We only included a single random effect for radiologist, as opposed to allowing two radiologist-specific effects that depend on disease status, because exploratory analyses fitting mixed-effects models (ignoring correlation within women) showed that the maximized log likelihoods were nearly equivalent for the model with

a single radiologist-specific effect and the model with two correlated random effects.

To test for endogeneity, we regressed each time-varying covariate (mammogram number, breast density, and disease status) on the previous outcome (mammogram result), adjusting for age and previous values of that covariate when appropriate. We found that the assumption of exogeneity does not hold for cancer status, because having a previous false positive mammogram was a significant predictor of future cancer (OR = 1.45, 95% CI = 1.18 to 1.76); There are a couple of possible explanations for this. First, some benign breast diseases picked up by mammography are predictive of future cancer. Second, this may be due to our definition of breast cancer, which only includes cases that are found within one-year and before the next screening examination. Some mammograms we considered to be false positive may have picked up cancer that was not diagnosed until after our follow-up period.

To account for endogeneity, we modelled the probability of having cancer at time t as a function of prior mammography result y_{it-1}at time $t-1$ using logistic regression:

$$\text{logit}\left[p\left(d_{it}=1\right)\right]=\zeta_{\mathbf{0}}+\zeta_1 y_{it-1}+\boldsymbol{\xi} H_{it-1}(\mathbf{x}).$$

where $H_{it-1}(\mathbf{x})$ includes prior age and prior density.

The prior distributions for $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\zeta}$and $\boldsymbol{\xi}$ were taken to be $normal(0,100)$ which is relatively flat across the range of typical logistic regression parameter values. The prior distribution for the precision τ for the radiologist-specific effect distribution was taken to be $gamma\left(2.1,2\right)$. The radiologist-specific effects are not expected to be greater than the typical size of logistic regression coefficients; thus, we chose values of the gamma distribution that put more weight on standard deviations less than 4. After updating the tuning

parameters for the Metropolis steps using the three-simulation strategy of Raftery and Lewis (1996), we ran three samplers for each model starting at dispersed values for 20,000 iterations each, throwing away the first 10,000 iterations for burn-in. Results are based on the 30,000 remaining iterations. To check convergence, the samplers were compared to verify convergence to the same posterior modes. For the Gauss-Hermite quadrature, 20 points were used.

For comparison, the marginal mean model was also fit using the three-step GEE strategy discussed in Section 3.

4.2 Results

The analyses include 123,083 screening mammograms on 73,216 women age 40-79, read by 41 radiologists. Among these women, 816 were diagnosed with breast cancer within the follow-up period. About half of the women had one observation (50.3%), 37.3% had two, 11.4% had three, and 1.0% had four or more observations over the 5 year time period. The number of mammograms read by each radiologist ranged from 109 to 10,287 with a median of 2,534, and the number of mammograms read for women diagnosed with breast cancer ranged from 0 to 73, with 40 radiologists seeing at least one woman subsequently diagnosed with cancer. Of the 816 women with cancer, 702 had a positive mammogram resulting in a crude sensitivity of 86.0% (Table 2). Of the 122,267 observations with no diagnosis of breast cancer, 107,696 had a negative mammogram giving a crude specificity of 88.1% (Table 2).

Table 2. *Number of observations (column percentages) for each mammogram result by breast cancer status.*

Mammogram Result	Breast Cancer	No Breast Cancer	Total N
Positive	702 (86.0%)	14,571 (11.9%)	15,273
Negative	114 (14.0%)	107,696 (88.1%)	107,810
Total N	816	122,267	123,083

The regression coefficients and widths of the confidence/credible intervals (CIs) from the marginalized multilevel models with and without correction for endogenous covariates, the three-step GEE approach, and the naïve (unadjusted) model are shown in Table 3. The Bayesian credible intervals are 95% highest posterior density intervals. In general, results are similar for all approaches. The most important predictor of a positive mammogram is having breast cancer. Having a first mammogram, increasing age, and increasing breast density are all associated with an increased probability of being recalled for further work-up of a mammogram; however, only breast density is significantly associated with poorer accuracy of mammography.

It is difficult to determine *a priori* if adjustment for correlation will result in smaller or larger variance estimates, since covariates vary both within and between clusters. However, comparing the CI widths reveals some clear patterns (Table 3). These patterns differ for the β coefficients (main effects), which are estimated from the entire study population, and the δ coefficients (interaction effects), which are effectively estimated only from women diagnosed with breast cancer. The GEE CIs for the β coefficients are wider than the naïve CIs, which is what we would generally expect since data are correlated within clusters. Except for the intercept, the β CIs from the marginalized multilevel model adjusting for endogeneity are narrower than the GEE CIs, hinting at efficiency gains for the likelihood-based approach.

Table 3. *Estimated regression coefficients and width of confidence/credible intervals (CI) from naive model, three-step GEE approach, and marginal multilevel model (MMM) without and with adjustment for endogenous covariates. Estimates with CIs that do not include zero are in bold. The naïve model assumes that all observations are independent.*

	Estimate			CI Width			
Parameter	Naïve/ GEE	MMM Exogenous	MMM Endogenous	Naïve	GEE	MMM Exogenous	MMM Endogenous
Intercept	-2.45	-2.34	-2.37	0.10	0.22	0.22	0.25
First screen	**0.48**	**0.52**	**0.51**	0.09	0.12	0.09	0.09
Age 40-49	**0.17**	**0.12**	**0.13**	0.12	0.14	0.12	0.11
Age 50-59	**0.22**	**0.21**	**0.21**	0.11	0.13	0.11	0.11
Age 60-69	**0.14**	**0.16**	**0.15**	0.12	0.14	0.12	0.12
Dense breasts	0.40	0.37	0.36	0.07	0.15	0.08	0.07
Breast cancer (BC)	**4.82**	**4.82**	**4.83**	1.00	0.96	1.03	1.00
First*BC	**0.39**	**0.42**	**0.46**	1.66	1.54	1.68	1.62
Age 40-49*BC	-0.69	-0.59	-0.55	1.28	1.11	1.25	1.24
Age 50-59*BC	-0.34	-0.35	-0.42	1.14	1.05	1.14	1.13
Age 60-69*BC	-0.02	-0.05	-0.04	1.21	0.94	1.29	1.20
Dense breasts*BC	**-1.18**	**-1.19**	**-1.16**	0.97	0.94	0.94	0.93
Previous FP		**0.70**	**0.70**			0.15	0.16
Previous FP*BC		1.39	0.88			3.03	3.21
Tau		7.06	6.26			4.80	4.35

FP=false positive, BC=breast cancer

The ratio of standard errors comparing the marginalized model to GEE range from 0.48 to 0.88. The CIs from the marginalized multilevel model with adjustment for endogeneity are of equal width or narrower than CIs from the unadjusted model.

For the δ coefficients, the GEE CIs are narrower than the naïve models. There are several possible explanations for this result. First, breast cancer status changes over time within women and varies both between and within radiologists. Second, this may be due to bias in standard error estimates resulting from the small number of clusters, with only 41 total radiologists (Mancl and DeRouen, 2001). The CIs from the marginalized multilevel model with adjustment for endogeneity are wider than the GEE CIs but narrower than the naïve CIs. The CIs from the marginalized multilevel model with adjustment for endogeneity are narrower than CIs from the unadjusted model.

Table 4. *Estimated sensitivity and specificity (95% confidence/credible intervals) by age, first versus subsequent mammography, and breast density, adjusted for other covariates in the model.*

	Sensitivity		Specificity	
	GEE	MMM	GEE	MMM
Overall	86.0 (83.4, 88.1)	86.5 (84.3, 89.1)	88.1 (87.1, 89.1)	87.3 (86.1, 88.5)
Mammogram number				
First	93.1 (87.3, 97.0)	94.9 (89.3, 97.8)	83.1 (81.8, 84.5)	81.8 (80.1, 83.3)
Subsequent	85.1 (82.3, 87.4)	85.7 (83.2, 88.4)	88.9 (87.8, 89.8)	88.2 (87.0, 89.3)
Age (years)				
40-49	80.4 (73.8, 86.1)	82.6 (75.7, 88.1)	88.0 (87.0, 88.9)	87.4 (86.1, 88.6)
50-59	85.9 (80.5, 90.4)	86.5 (82.4, 90.5)	87.4 (86.2, 88.5)	86.6 (85.2, 87.8)
60-69	88.5 (84.8, 91.7)	89.7 (85.1, 93.0)	88.2 (87.0, 89.4)	87.3 (85.9, 88.5)
70-79	87.3 (84.0, 90.1)	88.7 (83.5, 92.4)	89.6 (88.6, 90.5)	88.8 (87.6, 90.0)
Breast density				
Not Dense	91.2 (87.4, 94.3)	92.5 (88.9, 94.6)	90.2 (89.3, 91.2)	89.4 (88.3, 90.4)
Dense	82.8 (79.6, 85.5)	83.7 (80.1, 87.1)	86.2 (85.0, 87.4)	85.5 (84.1, 86.8)

Table 4 displays the accuracy measures from the GEE model and the marginalized multilevel model with adjustment for endogeneity, standardized to the overall distribution of the

Table 5. *Conditional sensitivity and specificity (95% credible intervals) by previous mammography result, marginalized over radiologist-specific effects.*

	Mode	(95% CI)
Sensitivity		
Previous TN	83.8	(80.9, 87.2)
Previous FP	97.9	(91.4, 99.8)
Specificity		
Previous TN	89.5	(88.5, 90.5)
Previous FP	81.3	(79.6, 82.9)

other covariates in the model among cancer cases for sensitivity and the distribution among non-cancer cases for specificity. Sensitivity is estimated to be slightly lower and specificity is slightly higher from the GEE model compared to the marginalized multilevel model. The estimated sensitivity of mammography is 86.0 (95% CI = 83.4 to 88.1) from the GEE model and 86.5% (95% CI = 84.3% to 89.1%) from the marginalized multilevel model. The estimated specificity is 88.1 (95% CI = 87.1 to 89.1) from the GEE model and 87.3% (95% CI = 86.1% to 88.5%) from the marginalized multilevel model. Sensitivity is higher and specificity is lower for first mammograms. Sensitivity and specificity are both lower for women with dense breasts.

The marginalized multilevel model provides additional, scientifically-interesting information on conditional accuracy given previous mammography result, displayed in Table 5. Results were marginalized over the radiologists-specific effects using Gauss-Hermite quadrature. Having a previous false positive mammogram is predictive of a future positive mammogram, independent of disease status (OR = 2.02, 95% CI = 1.86 to 2.18), but is not significantly associated with mammography accuracy (OR = 2.41, 95% CI = 0.72 to 17.8). Both the true positive and false positive rates are higher for women with a previous false positive

mammogram. Sensitivity is 83.8% (95% CI = 80.9% to 87.2%) for women with a previous true negative mammogram and 97.9% (95% CI = 91.4% to 99.8%) for women with a previous false positive mammogram. Specificity is 89.5% (95% CI = 88.5% to 90.5%) for women with a previous true negative mammogram and only 81.3% (95% CI = 79.6% to 82.9%) for women with a previous false positive.

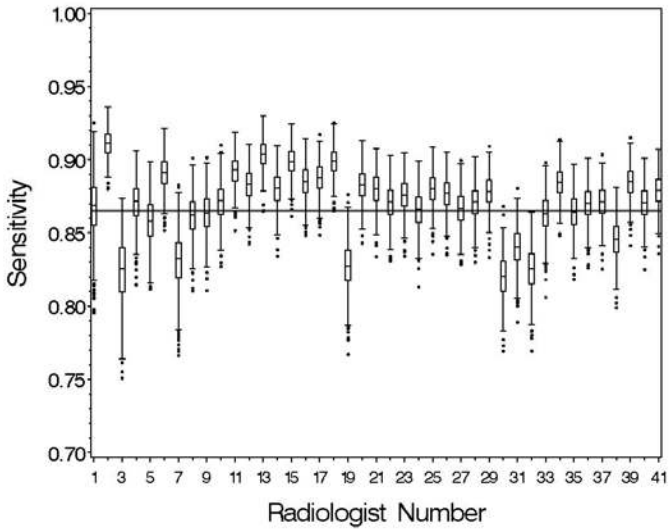


Figure 1: Posterior distribution of sensitivity by radiologist. Radiologists are ordered by increasing number of mammograms read where 1 corresponds to the radiologist that read the fewest mammograms (N=109) and 41 corresponds to the radiologist that read the most mammograms (N=10,287).

The posterior mode of the population precision for the radiologist-specific effects is 6.3. This corresponds to a standard deviation on the log-odds scale of 0.40. The posterior distributions for each radiologist's sensitivity and specificity, ordered by the total number of mammograms read, are shown in Figures 2 and 3. Posterior modes for sensitivity range from

82% to 91%. Specificity ranges from 80% to 91%.

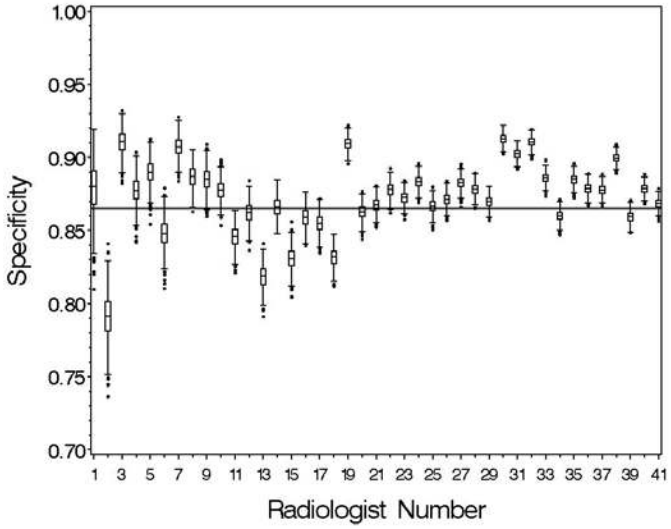


Figure 2: Posterior distribution of specificity by radiologist. Radiologists are ordered by increasing number of mammograms read where 1 corresponds to the radiologist that read the fewest mammograms (N=109) and 41 corresponds to the radiologist that read the most mammograms (N=10,287).

5 Discussion

This manuscript has focused on the development and comparison of two multilevel approaches for regression analysis of binary data. A GEE method that relies on a working independence assumption coupled with a three-step method for obtaining empirical standard errors is outlined. Likelihood-based methods implemented using Bayesian computational techniques are discussed, and implications of covariate endogeneity are addressed. Table 6 makes some broad comparisons of key advantages and disadvantages for each approach.

Table 6. *Comparison of key advantages and disadvantages of GEE and likelihood based approaches.*

Property	GEE	Likelihood-based
Number of clusters	(−) Requires a large number of clusters for valid empirical standard errors	(+) Permits general inference even with a small number of clusters
Missing data	(−) Requires missing completely at random (MCAR) or weighting inversely by non-missingness probability	(+) Requires either missing at random or MCAR
Estimation efficiency	(−) For non-nested clusters and endogenous covariates, working independence is required, which may be inefficient	(+) Optimal under correct model specification
Robustness	(+) Valid inference on regression parameters without requiring correct dependence model	(−) Valid inference requires correct specification of mean and dependence models
Computational ease	(+) Uses standard software with minor modification	(−) Requires tailored software
Time-varying covariates	(+) Consistent estimation of cross-sectional models by using working independence	(−) Requires exogeneity with appropriate lags or model for covariates

In our motivating application there are 41 top-level clusters, and GEE empirical standard error estimates may be negatively biased with such a modest number of clusters. Bayesian estimation is essentially exact, but with a small number of clusters can be sensitive to the variance component or regression parameter priors. Missingness issues were not considered important in our example, but without modification GEE may give biased results when data are not missing completely at random. Although efficiency may not seem an issue with over 120,000 observations, ultimately we have only 816 cancer cases, and only 114 false negative tests among these cases. Thus, our data contain substantial information regarding specificity but rather limited information regarding sensitivity. In prospective longitudinal studies, although many subjects are typically enrolled, the accrual of incident cases may be quite small for rare outcomes and thus efficient estimation can be crucial. Finally, in contrast to GEE, use of maximum likelihood or Bayesian methods requires correct dependence model specification for valid inference and tailored software to address the non-nested multilevel structure and endogenous covariates.

In this manuscript we develop two estimation methods for marginal regression inference that are valid under different assumptions about the distribution of the observed data. The ability to compare the primary regression results and assess whether conclusions are sensitive to the specific secondary dependence or missingness assumptions is valuable in practice. Ultimately, the choice of analysis method will depend on the particular characteristics of the application, including the plausibility of required assumptions and computational convenience.

6 Acknowledgements

This work was supported by the National Cancer Institute funded collaborative agreement U01CA86076 and by National Heart, Lung, and Blood Institute grant HL72966. The authors thank the Breast Cancer Surveillance Consortium for providing the data used in the example and Dr. Elizabeth Brown, an associate editor, and two reviewers for their helpful comments.

7 Appendix 1: MCMC algorithm

In the style of Gibbs Sampling, we first update the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in a single block, conditional on $\mathbf{u}$ and τ, using a Metropolis (random walk) step. Let $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T\right)^T$. A vector $\mathbf{z}$ is simulated from a multivariate normal distribution with mean zero and covariance matrix Σ and the candidate values $\boldsymbol{\theta}^*$ are taken to be the current values $\boldsymbol{\theta} + \mathbf{z}$. The tuning parameter Σ may be estimated using the three-simulation strategy of Raftery and Louis (1996), setting $\Sigma = \frac{2.3}{\sqrt{Q}}\Sigma^*$where Σ^* is the estimated conditional covariance matrix of $\boldsymbol{\theta}$ and Q is the length of $\boldsymbol{\theta}$. The acceptance probability for $\boldsymbol{\theta}^*$ is

$$\exp\left\{-\frac{\psi}{2}\sum_{q=1}^{Q}\left[\left(\theta_q^*\right)^2 - \theta_q^2\right]\right\}\prod_{i=1}^{N}\prod_{t=1}^{T_i}\frac{\left(\mu_{it}^C\left(\tau^*\right)\right)^{y_{it}}\left(1-\mu_{it}^C\left(\tau^*\right)\right)^{(1-y_{it})}}{\left(\mu_{it}^C\left(\tau\right)\right)^{y_{it}}\left(1-\mu_{it}^C\left(\tau\right)\right)^{(1-y_{it})}}.$$

We update the cluster-specific effects $\mathbf{u}$ conditional on $\tau, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ using a Metropolis (random walk) step. For each j, the candidate value u_j^* is set equal to the current value $u_j + z$ where z is a normal deviate with precision estimated as above. The acceptance probability

for the candidate value u_j^* is

$$\exp\left\{-\frac{\tau}{2}\left[(u_j)^2-u_j^2\right]\right\}\prod_{i=1}^{N}\prod_{t=1}^{T_i}\frac{\left(\mu_{it}^C(\tau^*)\right)^{y_{it}}\left(1-\mu_{it}^C(\tau^*)\right)^{(1-y_{it})}}{\left(\mu_{it}^C(\tau)\right)^{y_{it}}\left(1-\mu_{it}^C(\tau)\right)^{(1-y_{it})}}.$$

Similarly, to update the population precision τ we use a random walk step. Unlike standard hierarchical regression, the likelihood for the marginalized hierarchical model depends on τ after conditioning on $\mathbf{u}$ because β (and hence Δ) depends on τ. The acceptance probability for the candidate value $\tau^* = \tau + z$, where z is a normal deviate with precision estimated as above, is

$$\left(\frac{\tau^*}{\tau}\right)^{A-1+u/2}\exp\left(-B(\tau^*-\tau)-\frac{\tau^*-\tau}{2}\sum_{j=1}^{J}u_j^2\right)\prod_{i=1}^{N}\prod_{t=1}^{T_i}\frac{\left(\mu_{it}^C(\tau^*)\right)^{y_{it}}\left(1-\mu_{it}^C(\tau^*)\right)^{(1-y_{it})}}{\left(\mu_{it}^C(\tau)\right)^{y_{it}}\left(1-\mu_{it}^C(\tau)\right)^{(1-y_{it})}}.$$

The most time-consuming step is estimating Δ. To increase computational speed, τ may be updated along with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in a single block.

8 Appendix 2: Solving for Δ

If all covariates are exogenous, $\mu_{it-1}(z)$ and $\mu_{it-1}^M(H_{it}(\mathbf{x}))$ do not depend on $\mathbf{x}_{it}$ given $H_{it-1}(\mathbf{x})$. In this case, we can solve for Δ sequentially by first calculating Δ_{i1} and $\mu_{i1}(z)$ and then sequentially updating Δ_{it} and $\mu_{it}(z)$ given $\mu_{it-1}(z)$ and $\mu_{it-1}^M(H_{it}(\mathbf{x}))$. We can calculate Δ_{i1} using the Newton-Raphson algorithm to solve the equation that links the marginal mean and the conditional expectation:

$$\mu_{i1}^M = \int \text{logit}^{-1}(\Delta_{i1}+\sigma z)\,\phi(z)\,dz \tag{7}$$

where

$$\Delta_{i1}^{(n+1)} = \Delta_{i1}^{(n)} - \frac{\int \text{logit}^{-1}\left(\Delta_{i1} + \sigma z\right)\phi\left(z\right)dz - \mu_{i1}^{M}}{\int \text{logit}^{-1}\left(\Delta_{i1} + \sigma z\right)\left(1 - \text{logit}^{-1}\left(\Delta_{i1} + \sigma z\right)\right)\phi\left(z\right)dz}.$$

The integrals may be estimated using Gauss-Hermite quadrature. Given Δ_{i1}, $\mu_{i1}\left(z\right)$ =$\text{logit}^{-1}\left(\Delta_{i1} + \sigma z\right)$

We may then sequentially solve for $\mu_{it}\left(z\right)$ and Δ_{it} given $\mu_{it-1}\left(z\right)$ and $\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)$ as follows. First, we iteratively solve for Δ_{it} using the Newton-Raphson algorithm and Gauss-Hermite quadrature to solve the equation that links the marginal mean and the conditional expectation (3) where

$$\Delta_{it}^{(n+1)} = \Delta_{it}^{(n)} - \frac{\int h_{it1}\mu_{it-1}\left(z\right) + h_{it0}\left[1 - \mu_{it-1}\left(z\right)\right]\phi\left(z\right)dz - \mu_{it}^{M}}{\int h_{it1}\left[1 - h_{it1}\right]\mu_{it-1}\left(z\right) + h_{it0}\left[1 - h_{it0}\right]\left[1 - \mu_{it-1}\left(z\right)\right]\phi\left(z\right)dz},$$

and $h_{itk} = h\left(\Delta_{it}, k, z\right)$. Given Δ_{it} and $\mu_{it-1}\left(z\right)$, we may calculate

$$\mu_{it}\left(z\right) = h\left(\Delta_{it}, 1, z\right)\mu_{it-1}\left(z\right) + h\left(\Delta_{it}, 0, z\right)\left[1 - \mu_{it-1}\left(z\right)\right].$$

In the case of an endogenous covariate x_{itk}, we need to additionally sample values for the regression coefficients ζ and $\boldsymbol{\xi}$ from the GLM for x_{itk} and estimate $\mu_{it-1}\left(z\right)$ and $\mu_{it-1}^{M}\left(H_{it}(\mathbf{x})\right)$ using the factorization given in section 2, equation (4).

9 References

Agresti, A. , and Liu, I.-M. (1999). Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics* 55, 936-943.

Betensky, R. A., Talcott, J. A., Weeks, J.C. (2000). Binary data with two, non-nested sources of clustering, an analysis of physician recommendations for early prostate cancer treatment. *Biostatistics* 1, 219-230.

Bryk, A. and Raudenbush S. (1992). *Hierarchical Linear Models for Social and Behavioral Research, Applications and Data Analysis Methods*, Newbury Park, CA: Sage.

Carlin, J. B., Wolfe, R., Brown, C. H., and Gelman, A.(2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes.*Biostatistics*. 2, 397-416.

Daniels, M. J. and Gatsonis, C. (1999) Multilevel hierarchical generalized linear models with applications to health services research. *Journal of the American Statistical Association*. 94, 29-42.

Diggle, P.J. (1988). An approach to the analysis of repeated measures. *Biometrics*. 44, 959–971.

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

Gilks, W. R., Richardson, S., Spiegelhalter, D. J., (1996). *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall.

Goldstein, H. (1995), *Multilevel Models*, London: Arnold.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for lon-

gitudinal categorical data.*Biometrics 58, 342-351.*

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science.* 15(1):1-26.

Hedeker D., and Gibbons R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics.* 50, 933-944.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine.* 7, 305-315.

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

Lumley, T. and Mayer-Hamblett, N. (2003). Asymptotics for marginal generalized linear models with sparse correlations, *University of Washington Biostatistics Techinical Report.*

Mancl, L. A., and DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* 57(1): 126-134.

Mayer-Hamblett, N., Self, S. (2001). A regression modeling approach for describing patterns of HIV genetic variation. *Biometrics*, 57(2), 449-460.

Metropolis, N., Rosenbluth, A. W. , Rosenbluth, M. N. , Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics* 21, 1087-1091.

Pepe, M. S. (2003) *The statistical evaluation of medical tests for classification and prediction.* Oxford: Oxford University Press.

Pepe, M. S., and Anderson, G. A. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation* 23(4), 939-951.

Raftery, A. E. and Lewis, S. M. (1996). Implementing MCMC in *Markov Chain Monte Carlo in Practice*. (Eds: W. R. Gilks, S. Richardson, D. J. Spiegelhalter), London: Chapman and Hall. pp. 115-130.

Robins, J., Rotnitzky A., and Zhao L.-P. (1995).Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. ,*Journal of the American Statistical Association* 90, 106–121.

Robins, J., Greenland, S., and Hu F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association* 94, 687–712.

Rodriguez, G. and Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society A* 164: 339-355.

SAS Institute Inc. (2000). SAS OnlineDoc®, Version 8.

Shults, J. and Morrow, A. L. (2002). Use of quasi-least squares to adjust for two levels of correlation. *Biometrics* 58 (3), 521-530.

Wang, Y.-G., and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika* 90, 29-41.