University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2014 *Paper* 38

Regression modeling of longitudinal binary outcomes with outcome-dependent observation times

Kay See Tan* Andrea B. Troxel† Stephen E. Kimmel‡
Kevin G. Volpp** Benjamin French††

*University of Pennsylvania, kaystan@mail.med.upenn.edu
†University of Pennsylvania, atroxel@mail.med.upenn.edu
‡University of Pennsylvania, stevek@mail.med.upenn.edu
**University of Pennsylvania, volpp70@exchange.upenn.edu
††University of Pennsylvania, bcfrench@mail.med.upenn.edu

Regression modeling of longitudinal binary outcomes with outcome-dependent observation times

Kay See Tan, Andrea B. Troxel, Stephen E. Kimmel, Kevin G. Volpp, and Benjamin French

Abstract

Conventional longitudinal data analysis methods assume that outcomes are independent of the data-collection schedule. However, the independence assumption may be violated, for example, when adverse events trigger additional physician visits in between prescheduled follow-ups. Observation times may therefore be associated with outcome values, which may introduce bias when estimating the eect of covariates on outcomes using standard longitudinal regression methods. Existing semi-parametric methods that accommodate outcome-dependent observation times are limited to the analysis of continuous outcomes. We develop new methods for the analysis of binary outcomes, while retaining the exibility of semi-parametric models. Our methods are based on counting process approaches, rather than relying on possibly intractable likelihood-based or pseudo-likelihood-based approaches, and provide marginal, population-level inference. In simulations, we evaluate the statistical properties of our proposed methods. Comparisons are made to 'naive' GEE approaches that either do not account for outcome-dependent observation times or incorporate weights based on the observation-time process. We illustrate the utility of our proposed methods using data from a randomized controlled trial of interventions designed to improve adherence to warfarin therapy. We show that our method performs well in the presence of outcome-dependent observation times, and provide identical inference to 'naive' approaches when observation times are not associated with outcomes.

Regression modeling of longitudinal binary outcomes with outcome-dependent observation times

Kay See Tan[1,*], Andrea B Troxel[1], Stephen E Kimmel[1,2], Kevin G Volpp[2], and Benjamin French[1]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania, PA 19104, U.S.A.

[2]Department of Medicine, University of Pennsylvania Perelman School of Medicine, U.S.A.

**email:* kaystan@upenn.edu

SUMMARY: Conventional longitudinal data analysis methods assume that outcomes are independent of the data-collection schedule. However, the independence assumption may be violated, for example, when adverse events trigger additional physician visits in between prescheduled follow-ups. Observation times may therefore be associated with outcome values, which may introduce bias when estimating the effect of covariates on outcomes using standard longitudinal regression methods. Existing semi-parametric methods that accommodate outcome-dependent observation times are limited to the analysis of continuous outcomes. We develop new methods for the analysis of binary outcomes, while retaining the flexibility of semi-parametric models. Our methods are based on counting process approaches, rather than relying on possibly intractable likelihood-based or pseudo-likelihood-based approaches, and provide marginal, population-level inference. In simulations, we evaluate the statistical properties of our proposed methods. Comparisons are made to 'naïve' GEE approaches that either do not account for outcome-dependent observation times or incorporate weights based on the observation-time process. We illustrate the utility of our proposed methods using data from a randomized controlled trial of interventions designed to improve adherence to warfarin therapy. We show that our method performs well in the presence of outcome-dependent observation times, and provide identical inference to 'naïve' approaches when observation times are not associated with outcomes.

KEY WORDS: Informative observation times; joint models; observation-time process; outcome process; outcome-dependent follow-up; semi-parametric regression.

1. Introduction

Longitudinal studies typically focus on an explicit outcome of interest collected over time. The data-collection schedule may constitute an implicit outcome (Rizopoulos, 2012), in that the timing or frequency of data collection may communicate information regarding features of the study design or patient-level characteristics. Consider the use of warfarin, a commonly prescribed oral anticoagulant. A patient on warfarin requires frequent monitoring, based on the international normalized ratio (INR), due to the drug's narrow therapeutic range. Anticoagulation levels above or below the therapeutic range increase the risk of bleeding or thromboembolism, respectively (Hylek et al., 1996). An out-of-range INR typically triggers a dose change (Brigden et al., 1998); a physician may request multiple closely spaced follow-up visits to monitor the impact of the dose change on INR response (Figure 1). In such settings, the intensity of events such as follow-up visits may depend on previous outcomes and measured or unmeasured covariates. If interest lies in estimating the effect of observed covariates on the probability of being out of therapeutic range, then it is necessary to incorporate the data-collection schedule in the estimation procedure. We focus on a marginal mean regression model to estimate the association between observed covariates and a binary outcome of interest. We refer to the longitudinal outcomes as the outcome process and the occurrence of data collection over time as the observation-time process.

[Figure 1 about here.]

If the probability of having a follow-up visit depends upon previous outcomes and measured or unmeasured covariates, then the outcome and observation-time processes are dependent and conventional longitudinal data analysis methods such as generalized estimating equations (GEE) (Liang and Zeger, 1986) that ignore the observation-time process may provide biased estimates of covariate-outcome associations (Sun et al., 2005; French and Heagerty, 2009). We have introduced a framework to describe the potential relationship between the outcome

and observation-time processes based on assumptions regarding conditional independence (Tan et al., under review). Specifically, we assume that the outcome and observation-time processes are conditionally independent given past observed covariates in the outcome model, past observed covariates in the observation-time model, and/or shared, unobserved latent variables.

Various methods have been proposed to account for potential dependence between the observation-time process and longitudinal binary outcomes. Fitzmaurice et al. (2006) proposed a pseudolikelihood estimator utilizing a linear approximation of the conditional distribution of the binary outcomes. The estimator requires strong assumptions about the observation-time process (e.g., that the conditional distribution of the outcome process at time t is independent of the observation-time process given the most recent observed value of response prior to t) and does not allow for explicit specification of the observation-time model. Other authors have adopted an estimating equations approach, explicitly specifying the observation-time models to be incorporated into the estimation of the outcome model (Lin, Scharfstein, and Rosenheck, 2004; Bůžková and Lumley, 2007). Although these estimating equations approaches allow weaker assumptions about the observation-time process, they require a parametric structure for the mean trajectory of the outcomes over time.

Several authors have proposed semi-parametric estimation procedures that assume a non-parametric structure for the mean trajectory of the longitudinal outcomes and a parametric structure for covariate effects (Lin and Ying, 2001; Bůžková and Lumley, 2009; Liang, Lu, and Ying, 2009; Sun, Song, and Zhou, 2011). These models provide flexibility when the focus is on the effect of a particular covariate of interest, while the effect of time is considered a nuisance. Closed-form solutions for the mean trajectory and the parameters of interest are derived using the properties of mean-zero processes. These proposed semi-parametric

estimation procedures allow for more flexible modeling of the longitudinal outcomes, but are currently limited to continuous and count outcomes.

We consider a joint model approach to semi-parametric marginal regression to accommodate outcome-observation dependence in longitudinal studies with binary outcomes. Through the incorporation of observation-level visit-intensity weights and shared latent variables, our proposed joint model approach provides flexibiliy to accommodate the assumptions of conditional independence given observed covariates and/or subject-level latent variables, while not imposing a parametric assumption on the mean trajectory of the longitudinal binary outcomes.

In Section 2, we detail assumptions regarding conditional independence between the outcome and observation-time processes. In Section 3, we introduce a comprehensive estimation procedure for regression modeling of binary outcomes in the presence of outcome-dependent observation times. We present simulation studies to evaluate the performance of our proposed procedure under alternative outcome-observation dependence mechanisms in Section 4, and illustrate its application to data from a warfarin study in Section 5. Section 6 provides discussion and concluding remarks.

2. Model formulation and assumptions

We consider a longitudinal study with n independent subjects in the study interval $[0, \tau]$, for which τ is the maximum study duration. For subject i, $i = 1, \ldots, n$, let $Y_i(t)$ denote a binary outcome of interest at time t, and $X_i(t)$ denote a $p \times 1$ vector of possibly time-dependent covariates. Unless otherwise specified, we consider only external covariates, such that any time-dependent covariate process at time t is conditionally independent of all previous outcomes, given the history of the covariate process (Kalbfleisch and Prentice, 2002). $Y_i(\cdot)$ is measured at m_i observation times $0 \leq T_{i1} < T_{i2} < \cdots < T_{im_i} \leq \tau$, for which m_i denotes the number of follow-up measurements on the i^{th} individual. Using counting process

notation, let $N_i(t) = \sum_{s \leq t} dN_i(s)$ denote the number of observations on the i^{th} subject by time $t \leq C_i$, in which C_i is the censoring time. The indicator variable $dN_i(t)$ equals 1 if a follow-up visit occurred on the i^{th} individual at time t and equals 0 otherwise. We assume non-informative censoring, such that $\Pr[Y_i(t) = 1 \mid X_i(t), C_i \geq t] = \Pr[Y_i(t) = 1 \mid X_i(t)]$. That is, the covariate-outcome associations are the same in those who are censored at C_i as those who have survived beyond C_i.

2.1 *Semi-parametric outcome model*

We assume that primary scientific interest lies in a semi-parametric regression model for the longitudinal binary outcomes. We extend the semi-parametric linear regression model for continuous outcomes proposed by Lin and Ying (2001) to binary outcomes $Y_i(t)$ under independent or dependent observation times:

$$\Pr[Y_i(t) = 1 \mid X_i(t)] = \text{expit}\{\mu(t) + \beta' X_i(t)\} = \frac{\mu(t) + \beta' X_i(t)}{1 + \exp\{\mu(t) + \beta' X_i(t)\}}, \tag{1}$$

for which $\mu(t)$ is an arbitrary function of time and β is a $p \times 1$ vector of regression parameters of interest.

The semi-parametric outcome model assumes a parametric structure for the effect of $X_i(t)$ and a non-parametric structure for $\mu(t)$ (Lin and Carroll, 2001; Sun et al., 2005). Model (1) describes the marginal mean of $Y_i(\cdot)$ without specifying its correlation structure and distributional form; hence, it is appealing if the effects of $X_i(t)$ are of interest, but the effect of time is considered a nuisance.

The parameter β in model (1) represents the primary target of inference: the marginal association between a set of covariates and an outcome of interest among a population of individuals. Model (1) does not condition on the entire covariate process or on past outcomes. Instead, it conditions on covariate information available at time t, which may include baseline covariates, covariates measured at or before t, or summaries of the covariate history, i.e., a partly conditional mean regression model (Pepe and Couper, 1997).

2.2 *Observation-time model*

The observation-time process describes the timing and intensity of follow-up visits and is characterized by a standard recurrent events model. We introduce a non-negative latent variable η_i with mean 1 and unknown variance σ^2. Given observation-time model covariates $Z_i(t)$ and η_i, the recurrent event process $N_i(\cdot)$ is a non-homogeneous Poisson process with intensity function (Pepe and Cai, 1993; Lin et al., 2000):

$$\lambda_i(t) = \eta_i \lambda(t) \exp\{\gamma' Z_i(t)\}, \quad t \in [0, \tau] \tag{2}$$

for which γ is a vector of unknown parameters and $\lambda(t)$ is an arbitrary baseline intensity function with $\Lambda(t) = \int_0^t \lambda(u)du$. Model (2) implies that the occurrence of observations follows a proportional intensity model, in which η_i inflates or deflates the visit intensity. If the censoring time is independent of the observation-time process, then the parameter γ can be consistently estimated by $\hat{\gamma}$ from the following estimating function (Lin et al., 2000):

$$U(\gamma) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i(t) - \bar{Z}(t;\gamma)\} dN_i(t), \tag{3}$$

for which:

$$\bar{Z}(t;\gamma) = \frac{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' Z_i(t)\} Z_i(t)}{\sum_{i=1}^{n} \xi_i(t) \exp\{\gamma' Z_i(t)\}},$$

and $\xi_i(t) = I(C_i > t)$. The parameter γ provides information regarding the observation-time model, but is considered a nuisance because our interest is in estimating the association parameter β from the outcome model. However, incorporating the observation-time process into estimation of β in a joint model facilitates reliable estimation under outcome-observation dependence, which we detail in the next section.

2.3 *Assumptions regarding conditional independence*

We identify the source of dependence between the outcome and observation-time processes using one of three outcome-observation dependence mechanisms:

(M1) Conditional independence given past outcome-model covariates;

(M2) Conditional independence given past observation-time model covariates;

(M3) Conditional independence given shared latent variables.

For the remainder of the paper, conditional independence given covariates implies conditional independence given past observed covariates.

(M1) Conditional independence given outcome-model covariates

The first mechanism assumes that the outcome process is conditionally independent of the observation-time process given outcome-model covariates $X_i(t)$, or a subset of $X_i(t)$:

$$\mathrm{E}[dN_i(t) \mid X_i(t), Y_i(t), C_i \geq t] = \mathrm{E}[dN_i(t) \mid X_i(t)].$$

The probability of observation at time t depends on $X_i(t)$, $Y_i(t)$, and C_i only through outcome-model covariates $X_i(t)$, hence is plausible if the occurrence of a follow-up visit is due to the features of the study design, rather than subject-specific behaviors.

(M2) Conditional independence given observation-time model covariates

The second mechanism assumes that the occurrence of a follow-up visit depends on observation-time model covariates $Z_i(t)$:

$$\mathrm{E}[dN_i(t) \mid X_i(t), Z_i(t), Y_i(t), C_i \geq t] = \mathrm{E}[dN_i(t) \mid Z_i(t)].$$

The probability of observation at time t depends on $X_i(t), Z_i(t), Y_i(t)$, and C_i only through observation-time model covariates $Z_i(t)$. $Z_i(t)$ includes the full or partial subset of the outcome-model covariates and any additional measured covariates at or before time t, as well as previous outcomes. Note that (M1) $\subset$ (M2) because $X_i(t) \subset Z_i(t)$.

(M3) Conditional independence given shared latent variables

The third mechanism assumes that the outcome process is conditionally independent of the observation-time process given outcome-model covariates $X_i(t)$, and an unmeasured mean-one subject-specific latent variable η_i:

$$\mathrm{E}[dN_i(t) \mid X_i(t), Y_i(t), \eta_i, C_i \geq t] = \mathrm{E}[dN_i(t) \mid X_i(t), \eta_i].$$

η_i conveys information regarding subject-specific unmeasured confounders and propensity for physician visits.

Our outcome-observation dependence mechanisms provide a framework for reliable estimation of covariate-outcome associations. Mechanisms (M2) and (M3) allow the probability of an observation to depend on unmeasured patient characteristics in addition to measured observation-time model covariates, and hence place fewer restrictions on the probability of having a visit than (M1); these are reasonable assumptions in most observational studies. However, (M2) and (M3) may require more advanced analysis methods to provide valid inference, which we introduce in the following section.

3. Estimation and inference

In this section, we detail a new estimation procedure to estimate covariate-outcome associations with binary outcomes in a joint modeling approach under any combination of the three outcome-observation dependence mechanisms described in the previous section.

3.1 *Estimators*

3.1.1 *Estimator under M1.* Given the semi-parametric outcome model (1), and the observation-time model $\mathrm{E}[\,dN_i(t) \mid X_i(t)] = \exp\{\gamma' X_i(t)\}\, d\Lambda(t)$, we can define the zero-mean stochastic process for binary outcomes as:

$$M_i(t;\beta,\gamma) = \int_0^t \Big[Y_i(s)\xi_i(s)\, dN_i(s) - \mathrm{expit}\{\mu(s) + \beta' X_i(s)\}\xi_i(s) \exp\{\gamma' X_i(s)\}\, d\Lambda(s) \Big]. \quad (4)$$

$M_i(t;\beta,\gamma)$ is appropriate if it is assumed that the occurrence of a follow-up visit is a feature of the study design or known patient characteristics and not due to previous outcomes or unmeasured patient characteristics.

3.1.2 *Estimator under M2.* For continuous outcomes, Bůžková and Lumley (2009) proposed a method that relaxes the assumption of (M1) and accommodates (M2) by applying

observation-level weights to the estimating equation to account for dependence through covariates in the observation-time model, $Z_i(t)$. Recall that $Z_i(t)$ may include the outcome-model covariates $X_i(t)$ and summaries of past outcomes.

Given the marginal semi-parametric regression model (1), the observation-level weights standardize the observed data to the time-specific underlying population under the proportional rate model for observation times $\mathrm{E}[dN_i(t) \mid Z_i(t)] = \exp\{\gamma' Z_i(t)\} d\Lambda(t)$. One particular observation-level weight with variance-stabilizing properties is:

$$\rho_i(t; \gamma, \delta) = \frac{\exp\{\gamma' Z_i(t)\}}{\exp\{\delta' X_i(t)\}},$$

for which δ is estimated by $\hat{\delta}$ using (3) conditioning on $X_i(t)$. The zero-mean process $M_i(t; \beta, \gamma)$ from (4) can then be extended as:

$$M_{i1}(t; \beta, \gamma, \delta) = \int_0^t \frac{1}{\rho_i(s, \gamma, \delta)} \Big[Y_i(s) \xi_i(s)\, dN_i(s) - \mathrm{expit}\{\mu(s) + \beta' X_i(s)\} \xi_i(s) \exp\{\gamma' Z_i(s)\}\, d\Lambda(s) \Big]. \quad (5)$$

3.1.3 *Estimator under M2 and M3.* To allow outcome-observation dependence through observed covariates and unobserved latent variables, the outcome model (1) can be extended to:

$$\Pr[Y_i(t) = 1 \mid X_i(t)] = \mathrm{expit}\{\mu(t) + \beta' X_i(t) + \eta_{i1}' Q_i(t)\}, \quad (6)$$

in which $Q_i(t)$ is a $q \times 1$ subvector of $X_i(t)$ and η_{i1} is a q-dimensional vector of subject-specific latent variables that represent subject-level propensity for visit (Liang et al., 2009). The observation-time model can be expressed as:

$$\mathrm{E}[d\Lambda_i(t) \mid Z_i(t)] = \eta_{i2} \exp\{\gamma' Z_i(t)\}\, d\Lambda(t), \quad (7)$$

in which η_{i2} is a mean-one, non-negative latent variable. The distribution of η_{i2} may depend on observed time-independent outcome-model covariates V_i with $\mathrm{E}[\eta_{i2} \mid V_i] = 1$. Discussion regarding covariate-dependent latent variables or frailties can be found in recent literature (Heagerty and Kurland, 2001; Neuhaus and McCulloch, 2006; Liu et al., 2011; McCulloch

and Neuhaus, 2011). The latent variables from models (6) and (7) are assumed to be linearly linked through $\mathrm{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1)$. The parameter θ describes the association between the outcome and observation-time processes. Thus, to ensure that β retains a marginal interpretation with the inclusion of the latent variable, we define $B_i(t) = \mathrm{E}[(\eta_{i2} - 1) \mid m_i, C_i]Q_i(t)$ as a fixed covariate that incorporates the subject-specific propensity for visit. The outcome model (6) can be re-expressed as:

$$\Pr[Y_i(t) = 1 \mid X_i(t), B_i(t)] = \mathrm{expit}\{\mu(t) + \beta' X_i(t) + \theta' B_i(t)\}. \tag{8}$$

Next, we re-express the observation-time model. Let $\mathcal{Z}_i(t) = \{Z_i(s) : 0 \leq s < t\}$ denote the covariate history of Z_i up to t. Following the results from Huang et al. (2010), the event times $(t_{i1} < t_{i2} < \cdots < t_{im_i})$ of the i^{th} subject conditional on $\{C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)\}$ are order statistics of a set of independent and identically distributed random variables with the density function:

$$p\{t_{i1} < t_{i2} < \cdots < t_{im_i} \mid C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)\} = \frac{\exp\{\gamma' Z_i(t)\} d\Lambda(t)}{\int_0^{C_i} \exp\{\gamma' Z_i(s)\} d\Lambda(s)}.$$

Define $\pi(t; Z_i) = \int_0^t \exp\{\gamma' Z_i(s)\} d\Lambda(s)$. It follows that:

$$\mathrm{E}[dN_i(t) \mid C_i, m_i, \eta_{i2}, \mathcal{Z}(C_i)] = \xi_i(t) m_i \frac{d\pi(t; Z_i)}{\pi(C_i; Z_i)}.$$

Using both re-expressed outcome and observation-time models, the zero-mean process $M_{i1}(t; \beta, \gamma, \delta)$ from (5) can then be extended as:

$$M_{i2}(t; \beta, \theta, \gamma, \delta) = \int_0^t \frac{1}{\rho_i(s, \gamma, \delta)} \Big[Y_i(s)\xi_i(s)\, dN_i(s) - \mathrm{expit}\{\mu(s) + \beta' X_i(s) + \theta' \hat{B}_i(s)\}\xi_i(s) m_i \frac{d\pi(s, Z_i)}{\pi(C_i, Z_i)} \Big] \tag{9}$$

in the presence of both (M2) and (M3).

To estimate $B_i(t)$, we first estimate η_{i2} from the observation-time model. We utilize the property that given $\{\eta_{i2}, C_i, \mathcal{Z}(C_i)\}$, m_i follows a Poisson distribution with mean $\eta_{i2}\pi(C_i, Z_i)$ to obtain $\hat{\eta}_{i2} = \{\frac{m_i}{\pi(C_i, Z_i)} - 1\}$, so $\hat{B}_i(t) = \{\frac{m_i}{\pi(C_i, Z_i)} - 1\}Q_i(t)$. (9) is the most general formulation of the joint model and can accommodate (M1), (M2), and (M3); that is, provide valid

estimation of β under any combination of the three conditional independence mechanisms. Given a specific mechanism of (M1), (M2) or (M3), $M_{i2}(t;\beta,\theta,\gamma,\delta)$ can be reduced to (4) and (5). In subsequent sections, we proceed with estimation of β via the estimation equation (9), which we refer to as the 'proposed estimator.'

3.2 *Estimation procedure*

Unlike for continuous or count outcomes, there is no closed-form solution for $\mu(t)$ and β for binary outcomes based on the zero-mean process (9), by setting $M_{i2}(t;\beta,\theta,\gamma,\delta) = 0$. Computational issues may arise because $\mu(t)$ is infinite dimensional, and iterative procedures may be difficult with sparse data resulting from few subjects with visits at each unique observation time. To overcome these computational burdens, we impose a flexible structure on $\mu(t)$ using basis approximations. Generalizing the notation from Huang, Zhang, and Zhou (2007), suppose the smooth function $\mu(\cdot)$ can be approximated by a spline function such that $\mu(t) \approx \sum_{k=1}^{K_n} \varphi_k G_k(t) = \varphi' G(t)$ in which $\{G_k(\cdot), k = 1,\ldots,K_n\}$ is a basis system of B-splines, $\varphi = (\tau_1,\ldots,\tau_{K_n})'$ and $G(t) = (G_1(t),\ldots,G_{K_n}(t))'$. Let $\tilde{H}_i(t) = G_i(t)$ or $\tilde{H}_{ij} = G(T_{ij})$. (8) can thus be approximated by $\Pr[Y_i(t) = 1 \mid X_i(t), B_i(t)] = \text{expit}\{\varphi\tilde{H}_i(t) + \beta' X_i(t) + \theta' B_i(t)\}$. Let $s_1 < s_2 < \cdots < s_J$ denote the J distinct ordered observation times from all subjects $\{t_{ik}, i = 1,\ldots,n; k = 1,\ldots,m_i\}$. We propose to estimate β from (9) by the estimating equation:

$$\sum_{i=1}^{n}\sum_{k=1}^{m_i}\begin{pmatrix}\tilde{H}_i(t_{ik})\\ X_i(t_{ik})\\ \hat{B}_i(t_{ik})\end{pmatrix}\frac{Y_i(t_{ik})}{\rho_i(t_{ik},\gamma,\delta)}\xi_i(t_{ik})\,dN_i(t_{ik}) \tag{10}$$
$$-\sum_{j=1}^{J}\sum_{i=1}^{n}\begin{pmatrix}\tilde{H}_i(s_j)\\ X_i(s_j)\\ \hat{B}_i(s_j)\end{pmatrix}\frac{1}{\rho_i(s_j,\gamma,\delta)}\text{expit}\{\varphi'\tilde{H}_i(s_j) + \beta' X_i(s_j) + \theta'\hat{B}_i(s_j)\}\xi_i(s_j)m_i\frac{d\pi(s_j,Z_i)}{\pi(C_i,Z_i)} = 0.$$

γ and $\Lambda(t)$ can be estimated by $\hat{\gamma}$ from (3) and $\hat{\Lambda}(t) = \sum_{i=1}^{n}\int_0^t dN_i(s) \Big/ \sum_{j=1}^{n}\xi_j(s)\exp\{\gamma' Z_j(s)\}$. The number of equations represented by $\tilde{H}_i(\cdot)$ reflects K_n, the number of knots selected.

$\hat{\mu}(\cdot) = \hat{\varphi}'G(\cdot)$ estimates the non-parametric portion of the outcome model and $\hat{\beta}$ estimates the parametric portion of the outcome model, while $\hat{\theta}$ incorporates the effect of the visit process into the outcome model. Standard error estimation for our proposed estimation procedure can be obtained using a cluster bootstrap, in which subjects are sampled with replacement (Field and Welsh, 2007). Bootstrapping ensures that uncertainty from estimating μ and θ are accounted for in standard error estimate for $\hat{\beta}$.

3.3 *Parameter interpretation*

The inclusion of the fixed covariate $B_i(t)$ in (8) ensures that β retains a marginal interpretation. Consider a model with a binary treatment indicator X_{i1} and a confounder X_{i2}:

$$\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta \hat{B}_i\}, \tag{11}$$

such that $\hat{B}_i = (\hat{\eta}_{i2} - 1)Q_i(t)$ and β_1 is the parameter of interest. We examine four possible configurations of (11):

(i) $\Pr[Y_i(t) = 1 \mid X_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2}\}$;

(ii) $\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1)\}$;

(iii) $\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1)X_{i2}\}$;

(iv) $\Pr[Y_i(t) = 1 \mid X_i, \hat{B}_i] = \text{expit}\{\mu(t) + \beta_1 X_{i1} + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1)X_{i1}\}$;

In (i), corresponding to the outcome models in Sections 3.1.1 and 3.1.2, β_1 represents the difference in log odds of the response between two populations of treated and untreated individuals, regardless of their visit propensity. In (ii) and (iii), β_1 represents the difference in log odds of the response between two populations of treated and untreated individuals with the same value of X_{i2} and visit propensity. In (iv), the interpretation of β_1 is similar to the interpretation of the main effect in the presence of an interaction. The log odds for each treatment group can be expressed as:

$$\text{logit}\ \Pr[Y_i(t) = 1 \mid X_{i1} = 1, X_{i2}] = \beta_0 + \beta_1 + \beta_2 X_{i2} + \theta(\hat{\eta}_{i2} - 1);$$

$$\text{logit } \Pr[Y_i(t) = 1 \mid X_{i1} = 0, X_{i2}] = \beta_0 + \beta_2 X_{i2}.$$

Thus the comparison of the treatment groups results in the coefficient $\beta_1 + \theta(\hat{\eta}_{i2} - 1)$. Therefore, β_1 represents the difference in log odds of the response between two populations of treated and untreated individuals with the same value of X_{i2} and an average visit propensity (i.e., $\eta_{i2} = 0$).

4. Simulation study

We conducted simulation studies to evaluate the statistical properties of our proposed method under two outcome-observation dependence settings: (i) (M2) and (ii) (M2) and (M3). All simulations were conducted in R 2.13.1 (R Development Core Team, Vienna, Austria). For all simulations, we generated 1000 simulated datasets, each with $n = 100$ or 200 independent subjects. For comparison, we fit a GEE with a working independence correlation structure (IEE). We also fit an IEE that incorporated observation-level weights $\rho_i(t; \gamma, \delta)$ and $\hat{B}_i(t)$ as a covariate (weighted-IEE). All outcome models used B-splines with four degrees of freedom to approximate $\mu(t)$.

4.1 *Setting 1: Simulations under (M2)*

4.1.1 *Parameters.* In setting 1, we used covariates to induce correlation between the outcome and observation-time processes to satisfy (M2). We specified the outcome model as:

$$\Pr[Y_i(t) = 1 \mid X_i(t)] = \text{expit}\{\mu(t) + \beta_1 X_{i1}(t) + \beta_2 X_{i2}\}, \tag{12}$$

for which $\mu(t) = -1 + 0.5t^{-1/2}$, $\epsilon_i(t) \sim$ Normal(0,1), and (β_1, β_2) were the parameters of interest. The time-dependent covariate of interest $X_{i1}(t)$ took the form $X_{i1}\log(t)$, in which $X_{i1} \sim$ Uniform[0,1], and $X_{i2} \sim$ Bernoulli(0.5). We included an additional covariate X_{i3} drawn from a mixture distribution, for which $X_{i3} \sim$ Normal(2,1) if $X_{i1} \leq 0.5$ and $X_{i3} \sim$ Normal(0,4) if $X_{i1} > 0.5$.

Following the simulation procedure of Bůžková and Lumley (2009) based on a probit link approximation, we generated binary outcomes based on the following equation:

$$Y_i(t) = I\Big[f^*(t) + \beta_1^* X_{i1}(t) + \beta_2^* X_{i2} + \beta_3 X_{i3} + \phi_i + \epsilon_i(t) > 0\Big], \tag{13}$$

for which $f^*(t) = \mu(t)M - \beta_3 \mathrm{E}[X_{i3} \mid X_{i1}]$, $\beta_1^* = \beta_1 M$, $\beta_2^* = \beta_2 M$, and $M = \sqrt{\sigma_\epsilon^2 + \sigma_\phi^2 + \beta_3^2 \mathrm{Var}[X_{i3} \mid X_{i1}]}/1.7$. ϕ_i was a subject-specific latent variable that induced an exchangeable correlation structure on the outcomes from the same subject. We assumed ϕ_i was normally distributed with mean 0 and variance $\sigma_\phi^2 = 0.25$.

Model (13) describes the case when X_{i3} affects the covariate-outcome association by $X_{i1}(t)$. Proper marginalization over the additional covariate X_{i3}, the random effect, and the error term in (13) results in the marginal semi-parametric outcome model (12).

We generated observation times T_{ik} from a non-homogeneous Poisson process with intensity function $\lambda_i(t) = \eta_i \lambda(t) \exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_3 X_{i3}\}$, in which $\lambda(t) = \frac{\sqrt{t}}{2}$. Note that X_{i3} induced additional correlation between the outcome and observation-time processes, and X_{i3} was specified in the observation-time model but not in the marginal outcome model (12). The latent variable η_i was generated from a Gamma distribution with mean 1 and variance $\sigma_\eta^2 = 0.5$. The independent censoring time C_i was generated from Uniform[5,10]. To examine the performance of our proposed estimators under (M2), we considered various combinations of outcome parameters $\beta_1 = \log(1.5)$, $\beta_2 = \log(1.2)$, $\beta_3 = \{0, \log(0.5)\}$ and intensity parameters $\gamma_1 = 0.3$, $\gamma_2 = 0.2$, $\gamma_3 = (0, 0.2, 0.3)$. When $\gamma_3 = 0$, the outcome-observation dependence model satisfied (M1); when $\beta_3 \neq 0$ and $\gamma_3 \neq 0$, the outcome-observation dependence model satisfied (M2).

4.1.2 *Results.* Table 1 provides the estimated bias, empirical standard error estimates, and mean squared error estimates for estimation of β_1 in model (12) by the IEE, weighted-IEE, and our proposed method. If (M1) was satisfied ($\gamma_3 = 0$), i.e., the outcome and observation-time processes were conditionally independent given outcome-model covariates

$X_{i1}(t)$ and X_{i2}, then all three methods performed well. Biases in the estimates of β_1 were negligible. However, if (M1) was violated ($\gamma_3 \neq 0$ and $\beta_3 \neq 0$), i.e., the two processes had additional correlation induced by X_{i3}, then the bias under the proposed method was smaller than the bias under IEE. IEE estimates β_1 without accounting for the additional covariate X_{i3} in any manner, whereas the proposed method incorporates the effect of X_{i3} through observation-level weights. Thus, IEE provided biased estimates. The performance of the weighted-IEE was comparable to the proposed method; both methods provided comparable bias and mean squared errors of the covariate effects.

Because X_{i2} was independent of X_{i3}, the biases for β_2 were negligible under all three methods for all scenarios. This indicated that when the additional covariate is independent of an outcome-model covariate, the performance of IEE is comparable to the weighted-IEE and proposed methods.

[Table 1 about here.]

4.2 *Setting 2: Simulation under (M2) and (M3)*

4.2.1 *Parameters.* In setting 1, we focused on (M2). In setting 2, we examined the performance of our proposed estimator under (M3) when (M2) was satisfied. Following (6), the model of interest for binary outcomes in the presence of a latent variable representing visit propensity was:

$$\mathrm{P}[Y_i(t) = 1 \mid X_i(t)] = \mathrm{expit}\{\mu(t) + \beta_1 X_{i1}(t) + \beta_2 X_{i2} + \eta_{i1} Q_i(t)\}, \tag{14}$$

in which $\mu(t)$, $\epsilon_i(t)$, $X_{i1}(t)$ and X_{i2} were as defined in Section 4.1.1, and (β_1, β_2) were the parameters of interest. We introduced an additional covariate X_{i3}, defined as the mixture distribution as in Section 4.1.1, which affected the covariate-outcome association of $X_{i1}(t)$ through (M2). Extending (13), we generated data under both (M2) and (M3) with the following equation:

$$Y_i(t) = I\left[f^*(t) + \beta_1^* X_{i1}(t) + \beta_2^* X_{i2} + \beta_3 X_{i3} + \eta_{i1}^* Q_i + \epsilon_i(t) > 0\right], \tag{15}$$

for which $f^*(t)$, β_1^* and β_2^* were as defined in Section 4.1.1. With the inclusion of η_{i1}, we defined $\eta_{i1}^* = \eta_{i1} M$ and $M = \sqrt{\sigma_\epsilon^2 + Q_i^2\sigma_\phi^2 + \beta_2^2 \text{var}[X_{i2} \mid X_{i1}]}/1.7$.

The observation times T_{ik} were generated from a non-homogeneous Poisson process with intensity function $\lambda_i(t) = \eta_{i2}\lambda(t)\exp\{\gamma_1 X_{i1}(t) + \gamma_2 X_{i2} + \gamma_3 X_{i3}\}$, with $\lambda(t) = \frac{\sqrt{t}}{2}$. The independent censoring time C_i was generated from Uniform[5,10]. The latent variable η_{i1} was defined as $\text{E}[\eta_{i1} \mid \eta_{i2}] = \theta(\eta_{i2} - 1) + \phi_i$, for which $\phi_i \sim \text{Normal}(0, \sigma_\phi^2)$ and $\sigma_\phi^2 = 1$.

We generated the latent variable η_{i2} in the observation-time model under two scenarios:

(1) η_{i2} from Gamma distribution with mean 1 and variance 0.5; hereby $\eta_{i2}^{(1)}$.

(2) η_{i2} from a mixture distribution, following Uniform[0.5,1.5] if $X_{i2} = 1$ and Gamma distribution with mean 1 and variance 0.7 if $X_{i2} = 0$; hereby $\eta_{i2}^{(2)}$.

$\eta_{i2}^{(2)}$ would imply covariate-dependent latent variable, as introduced in Section 3.1.3.

The coefficients were defined as $(\beta_1, \beta_2, \beta_3) = \log(1.5, 1.2, 0.5)$, $(\gamma_1, \gamma_2, \gamma_3) = (0.3, 0.2, 0.3)$, and $\theta = 1$. $\theta \neq 0$ in model (14) introduced correlation between the outcome and the observation-time processes through latent variables.

We let $Q_i = 1$ or $Q_i = X_{i1}$. When $Q_i = 1$, the effect of the latent variable η_{i1} was not modified by any covariates in the outcomes process. When $Q_i = X_{i1}$, the effect of the latent variable η_{i1} was modified by the value of X_{i1}. By varying $Q_i = (1, X_{i1})$ and $\eta_{i2} = (\eta_{i2}^{(1)}, \eta_{i2}^{(2)})$, we considered different ways the latent variables induced a relationship between the outcome and observation-time processes.

4.2.2 *Results.* Table 2 provides the estimated bias, empirical standard error estimates, and mean squared error estimates for the estimation of β_1 and β_2 in model (14). From Section 4.1.2, the inclusion of X_{i3} in the observation-time model satisfied (M2) and induced

additional correlation and biases the covariate-outcome association of $X_{i1}(t)$. The inclusion of η_{i1} in the outcome model satisfied (M3).

We focus on the performance of the methods under various combinations of η_{i2} ($\eta_{i2}^{(1)}$ or $\eta_{i2}^{(2)}$) and Q_i (1 or X_i). If η_{i1} was unrelated to any of the outcome-model covariates, (i.e., $\eta_{i2}^{(1)}$ and $Q_i = 1$), then all three methods performed well for β_2, while IEE provided heavily biased estimates of β_1, consistent with the results from Section 4.1.2. Under $\eta_{i2}^{(2)}$, in which the distribution of η_{i2}, and hence value of η_{i1}, depended on the status of X_{i2}, IEE provided biased estimates of β_2, while the weighted-IEE and the proposed method provided unbiased estimates. If the effect of η_{i1} was modified by the value of X_{i1} (i.e., $Q_i = X_{i1}$), then the bias for β_1 under the weighted-IEE and proposed method was smaller than IEE. Under the (M1) or (M2)-only assumption ($\theta = 0$, $\eta_{i1} = 0$), the proposed method is expected to be less efficient than IEE because the estimation procedure attempts to estimate θ, which results in loss of efficiency.

[Table 2 about here.]

4.3 *Summary*

The preceding simulation results quantified the potential for bias in estimated covariate-outcome associations under various outcome-observation dependence mechanisms. Under (M1), all three methods performed well. Under (M2), only the weighted-IEE and our proposed method performed well. Under (M3) when (M2) was satisfied, both the weighted-IEE and our proposed method performed well in the presence of a latent variable representing visit propensity in the outcome model, especially when the latent variables were associated with outcome-model covariates either (i) if the distribution of the latent variables was covariate-dependent, or (ii) the effect of the latent variable was modified by an outcome-model covariate. In all simulations, the weighted-IEE and the proposed method were the most

reliable and provided estimates with negligible biases under any combination of outcome-observation dependence mechanisms.

5. Application

5.1 *Background*

In this section, we apply our proposed joint model approach to data from a randomized controlled trial among patients on warfarin therapy. The goal of the trial was to determine the effectiveness of interventions designed to increase adherence to therapy, and thus improve anticoagulation control (Kimmel et al., 2007). The study randomized 362 subjects into four treatment arms, which we are unable to reveal in this preliminary analysis. The study protocol specified monthly follow-up visits, at which INR was measured. Physicians also scheduled as-needed visits in between protocol-required visits based on the patient's INR response.

The outcome $Y_i(t)$ in the outcome model was binary: 1 if the INR was outside the therapeutic range (out-of-range) at time t, and 0 otherwise. The primary exposure was treatment assignment. Descriptive analyses (data not shown) revealed several baseline covariates that were imbalanced across the four treatment groups ($P < 0.2$), and were thus adjusted for in the outcome model: employment status (working, disabled, or retired/unemployed), baseline age, race, Medicare insurance, education, history of diabetes, target INR range, and sub-therapeutic INR at baseline.

We considered two outcome models:

Model 1: $\Pr[Y_i(t) = 1 \mid X_i, B_i] = \text{expit}\{\mu(t) + \beta' X_i + \theta B_i\}$

Model 2: $\Pr[Y_i(t) = 1 \mid X_i, B_i] = \text{expit}\{\mu(t) + \beta' X_i + \theta_1 B_{i,\text{disabled}} + \theta_2 B_{i,\text{retired/unemployed}}\}$.

Recall that B_i includes the latent variable from the observation-time model, and $B_{i,Q_i}=(\hat{\eta}_{i2}-1)Q_i$. Model 1 assumed that the effect of the latent variable was not modified by any of the

outcome-model covariates. Model 2 assumed that the effect of subject-specific latent variables was different based on employment status.

The observation-time model was defined as: $\mathrm{E}[d\Lambda_i(t) \mid Z_i(t)] = \eta_{i2} \exp\{\gamma' Z_i(t)\} d\Lambda(t)$, in which $Z_i(t)$ included whether the INR was out-of-range at the previous visit. Thus, the observation-time model mirrored the clinical management of patients with suboptimal anticoagulation status who required additional follow-up. Outcome-model covariates were also screened for inclusion in the observation-time model. Univariable recurrent event models were used to assess unadjusted covariate associations with the observation times.

Censoring time C_i was defined as the time of the last follow-up visit at the study site. To estimate 95% confidence intervals (CI), we performed a cluster bootstrap in which subjects were sampled with replacement. The sampling procedure was repeated 1000 times and the 95% bootstrap CI was obtained from the 2.5^{th} and 97.5^{th} percentile of the empirical distribution produced from these 1000 estimates of β. The β estimates and 95% CI from the outcome models were exponentiated to obtain the odds ratios and corresponding 95% CI. Odds ratios less than 1 indicated decreased odds of out-of-range INR. For comparison, we fit a GEE with a working independence correlation structure (IEE) and an IEE that incorporated observation-level weights $\rho_i(t; \hat{\gamma}, \hat{\delta})$ and $\hat{B}_i(t)$ as a covariate (weighted-IEE). All outcome models used B-splines with 4 degrees of freedom to approximate $\mu(t)$.

5.2 *Results*

The estimates of γ from the observation-time model indicated that employment status was significantly associated with the observation times (Table 3). Patients who were disabled or retired/unemployed were more likely to have a visit compared to patients who were working. The median number of visits for those in the 'working' group was 6 (range, 3–11), while the median number of visits in the 'disabled' and 'retired/unemployed' groups were 8 (range, 2–16) and 7 (range, 1–24), respectively. Employment status may be a proxy for other

factors such as access to care and availability of time for physician visits. Patients were also significantly more likely to have a visit if the INR was out-of-range at the previous visit [$\hat{\gamma}$, 0.40; 95% CI:(0.37, 0.49)]. We found no significant interaction between employment status and out-of-range INR. The observation-level weights applied to the estimation procedures had a median of 1.21 and ranged from 0.88 to 1.71.

[Table 3 about here.]

The odds ratios (ORs) and 95% CIs from the outcome models are presented in Table 4. Under Model 1, the estimate of B_i was positive, implying that the outcome and observation-time processes were positively associated, such that patients with greater odds of being out-of-range (i.e., poorer anticoagulation status) had more frequent visits. With both observation-level weights and a latent variable, the OR estimates from weighted-IEE and the proposed method shifted toward the null for those disabled and retired/unemployed. Here, β for each employment status represents the difference in log odds of an out-of-range INR between populations of 'disabled' or 'retired/unemployed' individuals and 'employed' individuals with the same visit propensity.

[Figure 2 about here.]

Model 2 investigated whether the latent variable was associated with employment status. Based on the empirical distribution of η_{i2} by employment status (Figure 2), there was little evidence that the distribution of η_{i2} from the observation-time model differed by employment status. The OR estimates for those in the 'disabled' group were similar between Model 1 and 2, but the estimates for those in the 'retired/unemployed' group further attenuated, even though the confidence intervals were wide. Both the 'disabled' and the 'retired/unemployed' groups had more frequent visits compared to the working group, hence the incorporation of the observation-level weights and effects of latent variables based on employment status adjusted for potential outcome-observation dependence. Here, β for each

employment status represents the difference in log odds of an out-of-range INR between populations of 'disabled' or 'retired/unemployed' individuals and 'employed' individuals, all with average visit propensity.

[Table 4 about here.]

6. Discussion

In this paper, we presented a new approach to analyze longitudinal binary outcomes in the presence of outcome-dependent observation times. We introduced three mechanisms to describe the dependence between the outcome and observation-time processes, and showed that our proposed method is applicable under any combination of the mechanisms. Our proposed method performed as well as the weighted-IEE that incorporates observation-level weights and latent variables. Both methods performed better than the naïve IEE when the dependence between outcomes and observation times is parameterized using observation-time model covariates and/or latent variables.

The advantage of our proposed method over the weighted-IEE would be apparent in the case of more complicated data-collection schedules, such as when the censoring times are not independent of the outcome and observation-time processes. In addition, our proposed method allows explicit specification of separate models for the outcome and the observation-time processes. Although not our primary target of inference, the parameters in the observation-time model provide relevant information to clinicians regarding the timing of care provided to patients. The ability of our proposed method to explicitly specify the secondary model for the observation-time process can be extended to accommodate more complex data-collection schedules. For example, discontinuous risk intervals (i.e., patients may be in or out of the risk set based on disease status or treatment washout periods) may be incorporated in the observation-time model and therefore warrants future research.

Several key features of our approach are worth noting. First, we applied our proposed method to the analysis of binary outcomes, but our approach can be extended to other types of outcomes given an appropriate link function, such as the generalized logit link for a multinomial outcome. Second, we modeled the effect of time with B-splines instead of assuming a parametric structure. The potential gain in computational ease from the smooth spline approximation of $\mu(t)$ is countered by the potential loss in efficiency of estimation of the parameters of interest. Third, the validity of the proposed estimator is contingent upon correct specification of the observation-time model. One could utilize a Wald test for the importance of the additional covariates in $Z_i(t)$ to guide model building. Fourth, the model is able to accommodate censoring times that are dependent on the outcome and observation-time processes by estimating γ following the procedure in Huang et al. (2010). Finally, we note that in our application, patients were nested within physicians who made scheduling decisions based on the patient's anticoagulation status. Therefore, in addition to unmeasured patient characteristics, we may need to account for physician-level characteristics. Incorporating multiple sources of correlation, such as in a multi-level model, warrants future research.

Acknowledgements

We gratefully acknowledge the University of Pennsylvania for supporting this research.

References

Brigden, M. L., Kay, C., Le, a., Graydon, C., and McLeod, B. (1998). Audit of the frequency and clinical response to excessive oral anticoagulation in an out-patient population. *American journal of hematology* **59,** 22–7.

Bůžková, P. and Lumley, T. (2007). Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Canadian Journal of Statistics* **35,** 485–500.

Bůžková, P. and Lumley, T. (2009). Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Statistics in Medicine* **28,** 987–1003.

Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69,** 369–390.

Fitzmaurice, G. M., Lipsitz, S. R., Ibrahim, J. G., Gelber, R., and Lipshultz, S. (2006). Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. *Biostatistics* **7,** 469–485.

French, B. and Heagerty, P. J. (2009). Marginal mark regression analysis of recurrent marked point process data. *Biometrics* **65,** 415–422.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88,** 973–985.

Huang, C.-Y., Qin, J., and Wang, M.-C. (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* **66,** 39–49.

Huang, J. Z., Zhang, L., and Zhou, L. (2007). Efficient Estimation in Marginal Partially Linear Models for Longitudinal/Clustered Data Using Splines. *Scandinavian Journal of Statistics* **34,** 451–477.

Hylek, E. M., Skate, S. J., Sheehan, M. A., and Singer, D. E. (1996). An analysis of the lowest effective intensity of prophylactic anticoagulation for patients with nonrheumatic atrial fibrillation. *The New England Journal of Medicine* **335,** 540–546.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. Wiley, New York, 2 edition.

Kimmel, S. E., Chen, Z., Price, M., Parker, C. S., Newcomb, C. W., Samaha, F. F., and Gross, R. (2007). The Influence of Patient Adherence on Anticoagulation Control With Warfarin: Results from the International Normalized Ratio Adherence and Genetics (IN-RANGE) Study. *Archives of Internal Medicine* **167,** 229–235.

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

Liang, Y., Lu, W., and Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* **65,** 377–384.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62,** 711–730.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **96,** 103–126.

Lin, H., Scharfstein, D. O., and Rosenheck, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66,** 791–813.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96,** 1045–1056.

Liu, D., Kalbfleisch, J. D., and Schaubel, D. E. (2011). A positive stable frailty model for clustered failure time data with covariate-dependent frailty. *Biometrics* **67,** 8–17.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science* **26,** 388–402.

Neuhaus, J. M. and McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 859–872.

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88,** 811–820.

Pepe, M. S. and Couper, D. (1997). Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association* **92,** 991–998.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R.* Chapman and Hall/CRC Biostatistics Series. Chapman and Hall/CRC.

Sun, J., Park, D.-H., Sun, L., and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association* **100,** 882–889.

Sun, L., Song, X., and Zhou, J. (2011). Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times. *Journal of Statistical Planning and Inference* **141,** 2902–2919.

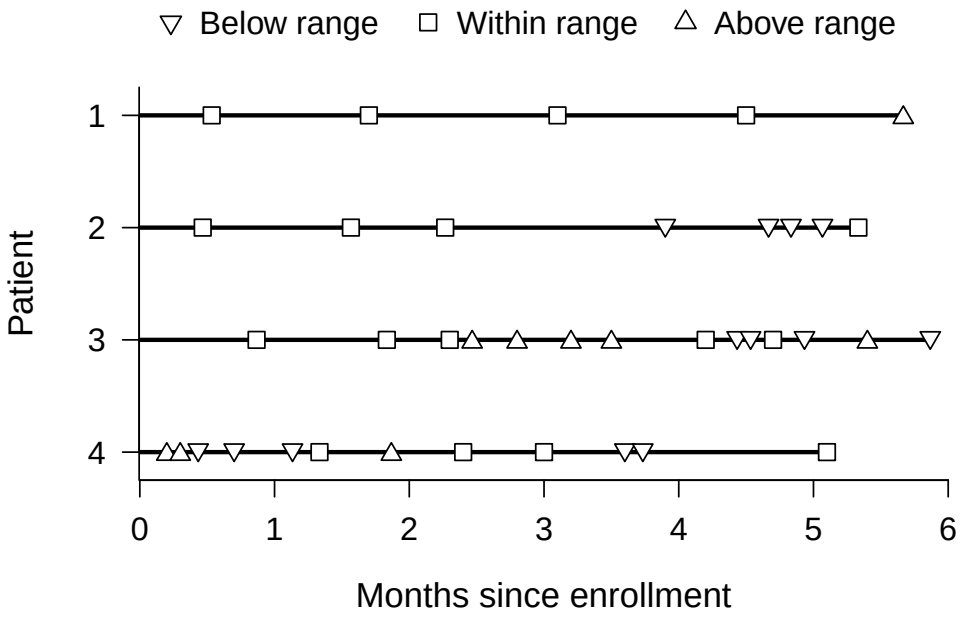

Figure 1: Observation times for four selected patients on warfarin and the corresponding observed outcomes: INR below, within, or above the therapeutic range.

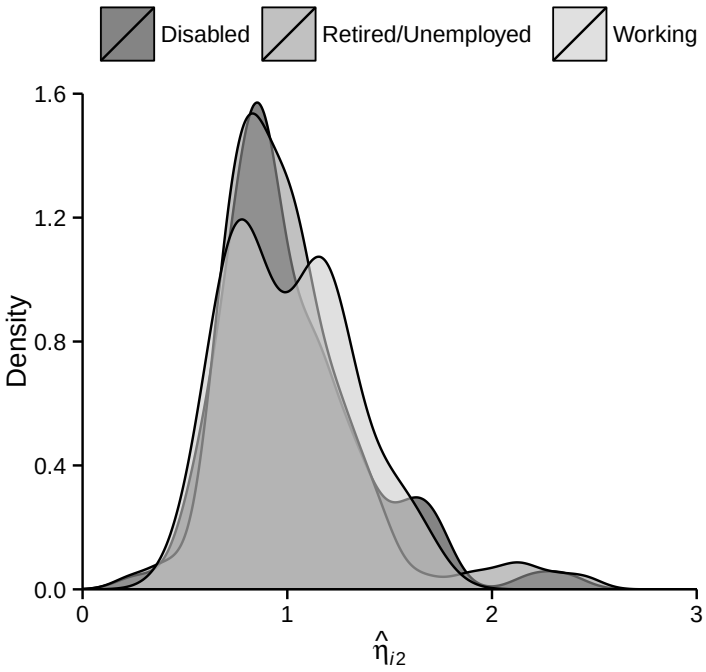

Figure 2: The empirical distribution of η_{i2} by employment status.

Table 1: Simulation results for $\beta_1 = \log(1.5)$: Bias, $\hat{\beta}_1 - \beta_1$; ESE, empirical sample error; MSE, mean squared error

β_3	n	γ_3	**IEE**			**Weighted-IEE**			**Proposed method**		
			Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
0	100	0	−0.02	0.25	0.06	−0.02	0.24	0.06	−0.01	0.26	0.07
		0.2	−0.04	0.24	0.06	−0.05	0.23	0.05	−0.04	0.24	0.06
		0.3	−0.04	0.24	0.06	−0.05	0.24	0.06	−0.05	0.27	0.08
	200	0	−0.02	0.18	0.03	−0.02	0.17	0.03	−0.02	0.18	0.03
		0.2	−0.04	0.18	0.03	−0.05	0.17	0.03	−0.05	0.18	0.03
		0.3	−0.04	0.17	0.03	−0.05	0.16	0.03	−0.05	0.20	0.04
log(0.5)	100	0	0.01	0.38	0.14	0.01	0.37	0.14	0.02	0.38	0.14
		0.2	−0.32	0.38	0.25	−0.06	0.36	0.13	−0.06	0.37	0.14
		0.3	−0.42	0.39	0.33	−0.04	0.36	0.13	−0.04	0.39	0.15
	200	0	−0.02	0.25	0.06	−0.01	0.25	0.06	−0.01	0.25	0.06
		0.2	−0.33	0.26	0.18	−0.07	0.25	0.07	−0.07	0.26	0.07
		0.3	−0.45	0.27	0.28	−0.06	0.25	0.07	−0.06	0.28	0.08

* All outcome models were fitted with B-splines with 4 degrees of freedom.

Table 2: Simulation results for $\beta_1 = \log(1.5)$, $\beta_2 = \log(1.2)$, $\theta = 1$: Bias, $\hat{\beta} - \beta$; ESE, empirical sample error; MSE, mean squared error

n	η_2	Q_i		**IEE**			**Weighted-IEE**			**Proposed method**		
				Bias	ESE	MSE	Bias	ESE	MSE	Bias	ESE	MSE
100	$\eta_2^{(1)}$	1	β_1	−0.40	0.47	0.38	−0.05	0.42	0.18	−0.05	0.44	0.20
			β_2	−0.01	0.46	0.21	0.03	0.44	0.19	0.03	0.44	0.20
		X_{i1}	β_1	−0.11	0.41	0.18	−0.05	0.34	0.12	−0.05	0.39	0.15
			β_2	0.00	0.38	0.15	0.02	0.34	0.12	0.02	0.35	0.12
	$\eta_2^{(2)}$	1	β_1	−0.42	0.43	0.36	−0.08	0.39	0.16	−0.08	0.41	0.17
			β_2	−0.39	0.41	0.32	−0.09	0.38	0.15	−0.09	0.39	0.16
		X_{i1}	β_1	−0.24	0.38	0.20	−0.09	0.33	0.12	−0.09	0.36	0.14
			β_2	−0.19	0.34	0.15	−0.03	0.32	0.10	−0.03	0.33	0.11
200	$\eta_2^{(1)}$	1	β_1	−0.42	0.33	0.29	−0.06	0.30	0.09	−0.05	0.30	0.10
			β_2	−0.02	0.31	0.10	0.02	0.30	0.09	0.02	0.30	0.09
		X_{i1}	β_1	−0.12	0.29	0.10	−0.06	0.24	0.06	−0.06	0.26	0.07
			β_2	0.00	0.27	0.07	0.02	0.24	0.06	0.02	0.24	0.06
	$\eta_2^{(2)}$	1	β_1	−0.42	0.30	0.27	−0.07	0.27	0.08	−0.07	0.27	0.08
			β_2	−0.40	0.29	0.25	−0.08	0.26	0.08	−0.08	0.26	0.08
		X_{i1}	β_1	−0.24	0.27	0.13	−0.08	0.23	0.06	−0.08	0.24	0.06
			β_2	−0.20	0.24	0.10	−0.02	0.22	0.05	−0.02	0.22	0.05

* All outcome models were fitted with B-splines with 4 degrees of freedom.
[a] Latent variable distributions: $\eta_{i2}^{(1)} : \eta_{i2} \sim \text{Gamma(mean=1}, \sigma^2 = 0.5)$;
$\eta_{i2}^{(2)} : \eta_{i2} \sim I[X_2 = 1]\text{Uniform}[0.5, 1.5] + I[X_2 = 0]\text{Gamma}(1, 0.5)$

Table 3: Parameter estimates and 95% CI of γ from the observation-time model

	$\hat{\gamma}$ (95% CI)
Employment Status	
Working	–
Disabled	0.13 (0.04, 0.29)
Retired/Unemployed	0.09 (0.00, 0.27)
Out-of-range INR at previous visit	0.40 (0.37, 0.49)

Table 4: Odds ratios (OR) and 95% confidence intervals (CI) for out-of-range INR

	Model 1[1]						**Model 2**[2]			
	IEE		**Weighted-IEE**		**Proposed method**		**Weighted-IEE**		**Proposed method**	
	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$	OR (95% CI)	$P^{\dagger}$
Employment Status		0.20		0.31		0.30		0.37		0.36
Working	–		–		–		–		–	
Disabled	1.34 (1.00, 2.12)		1.27 (0.92, 1.98)		1.28 (0.93, 2.00)		1.27 (0.91, 1.95)		1.28 (0.92, 2.03)	
Retired/Unemployed	1.22 (0.93, 2.20)		1.16 (0.85, 1.97)		1.17 (0.84, 2.02)		1.11 (0.80, 1.86)		1.12 (0.80, 1.89)	
θ	–		1.17 (0.98, 1.73)		1.15 (0.95, 1.80)		–		–	
θ_1	–		–		–		1.00 (0.58, 1.62)		0.98 (0.57, 1.56)	
θ_2	–		–		–		1.29 (1.05, 2.49)		1.26 (1.02, 2.70)	

* All outcome models were fitted with B-splines with 4 degrees of freedom.

† P-value corresponds to 2 degrees of freedom multivariate Wald test for β_{disabled}=$\beta_{\text{retired/unemployed}}$=0.

Outcome models:

[1] Model 1: $\text{P}[Y_i(t) = 1 \mid X_i] = \text{expit}\{\mu(t) + \beta' X_i + \theta B_i\}$

[2] Model 2: $\text{P}[Y_i(t) = 1 \mid X_i] = \text{expit}\{\mu(t) + \beta' X_i + \theta_1 B_{i,\text{disabled}} + \theta_2 B_{i,\text{retired/unemployed}}\}$.