Published in final edited form as:

Commun Stat Theory Methods. 2019 ; 48(12): 2982–3004. doi:10.1080/03610926.2018.1473599.

A Likelihood Based Approach for Joint Modeling of Longitudinal Trajectories and Informative Censoring Process

Miran A Jaffa^{*,1}, Ayad A Jaffa^{2,3}

¹Epidemiology and Population Health Department, Faculty of Health Sciences, American University of Beirut, Beirut, Lebanon

²Department of Biochemistry and Molecular Genetics, Faculty of Medicine, American University of Beirut, Beirut, Lebanon

³Department of Medicine, Medical University of South Carolina, Charleston, SC 29425, USA

Abstract

We propose a joint modeling likelihood-based approach for studies with repeated measures and informative right censoring. Joint modeling of longitudinal and survival data are common approaches but could result in biased estimates if proportionality of hazards is violated. To overcome this issue, and given that the exact time of dropout is typically unknown, we modeled the censoring time as the number of follow-up visits and extended it to be dependent on selected covariates. Longitudinal trajectories for each subject were modeled to provide insight into disease progression and incorporated with the number follow-up visits in one likelihood function.

Keywords

Biomarkers of kidney disease; Informative right censoring; Joint modeling; Latent random variables; Likelihood-based approach; Longitudinal data; Maximum likelihood estimation; Shared random effects

1. Introduction

In Longitudinal studies spanning over a long period of time, the predicament of missing data is inevitable. This issue gets more complicated when the mechanism that generates the incompleteness is dependent on the individual's characteristic and rate of change in the study variable. This phenomenon is referred to in longitudinal studies as informative right censoring (Wu and Bailey 1988). Specifically, the censoring process is used to refer to the mechanism that leads to incompleteness that is due to dropout before the termination of the study. If this mechanism is driven by or contains information on parameters of interest such as rate of change then it is considered informative censoring. For example, in a cohort of type 1 diabetic patients, it was shown that albumin excretion rate (AER) exceeding 300 mg/24 referred to as macroalbuminuria is associated with progression to stage 3 chronic

*Corresponding author: Miran A. Jaffa, Ph.D., Epidemiology and Population Health Department, Faculty of Health Sciences, American University of Beirut, P.O. Box 11-0236 Riad El-Solh, Beirut, Lebanon 1107 2020. ms148@aub.edu.lb.

The authors have no conflict of interest to declare and all authors read and approved the paper.

kidney disease that ultimately leads to informative right censoring due to renal failure (Molitch et al. 2010). The censoring process could be dependent on clinical factors and biomarkers such as plasma prekallikrein (PK) that need to be accounted for and incorporated in the dropout mechanism. Plasma PK is a serine protease synthesized in the liver and secreted into the plasma and is a risk marker for diabetic complications (Jaffa 2003; Jaffa 2016).

Conventional methods that do not account for informative right censoring are known to result in biased estimates (Tsiatis and Davidian 2001). Valid inferences require certain understanding of the link between the censoring mechanism and the longitudinal process and necessitate incorporation of the interdependency between the probability of censoring and changes in the longitudinal profiles. To link the survival and longitudinal processes joint modeling approaches were applied, and the main framework was based on the idea of having linear (mixed) models for the longitudinal process and joining it with the survival component using shared random effect (DeGruttola and Tu 1994; Tsiatis et al. 1995; Wulfoshn and Tsiatis 1997; Tsiatis and Davidian 2001, 2004; Guo and Carlin 2004; Vonesh, Greene, and Schluchter 2006; Hsieh, Tseng, and Wang 2006). In these approaches the longitudinal process was based on subject-specific random intercept and slope. More complex models were also proposed by having the longitudinal process follow smooth trajectories including splines and some mean-zero stochastic process that allow trends to vary with time (Taylor 1994; Henderson et al. 2000; Wang and Taylor 2001; Xu and Zeger 2001). Other models included the work of Ding and Wang 2008; Nathoo and Dean 2008; Rizopoulos, Verbeke, Lesaffre 2009; Wu, Liu, and Hu 2010; Jacqmin-Gadda et al 2010; Pan and Yi 2011). Most of these approaches were built on the concept of having mixed-effects longitudinal models and Cox model for the survival components and both parts are joined by shared random effects and the likelihood function is maximized using EM algorithms. A major challenge with joint modeling is its computational intensity and infeasibility especially in the presence of missing data (Wu, Liu and Hu 2010). Semiparametric models were another alternative for joint modeling that do not assume normality of random effects. These models either considered a class of smooth densities for these random effects or did not assume any parametric dependence between the censoring process and longitudinal outcomes (Tsiatis and Davidian 2001; Gallant and Nychka 1987; Yao, Wei, and Hogan 1998; Zhang et al. 1998; Glidden and Wei 1999; Lin and Ying 2001; Song, Davidian, and Tsiatis 2002; Song and Wang 2008). Convergence issues may arise under these models especially in the presence of missing data. Bayesian approaches to joint modeling of time-to-event and longitudinal process were also considered and implemented using the Markov chain Monte Carlo (MCMC) technique or semiparametric Bayesian joint modeling (Wang and Taylor 2001; Xu and Zeger 2001; Faucett and Thomas 1996; Brown and Ibrahim 2003; Brown and Ibrahim 2003). These approaches are characterized by inbuilt assumptions, the specifications of which are scarce (Tsiatis and Davidian 2001) and the impact of the violation of these assumptions on the inferences is not well determined. A review by Tsiatis and Davidian (2001) outlined the methodological advance and limitations of joint modeling inherent in computational complexity in maximizing the likelihood function and the unavailability of precise underlying assumptions.

Most of these methods did not account for informative right censoring or dropout. In this regard, several approaches were developed (Wang, Qin, Chiang 2001; Liu, Wolfe, Huang 2004; Rondeau et al 2007; Ye, Kalbfleisch, Schaubel 2007; Huang and Liu 2007; Liu, Huang, O'Quigley 2008; Liu and Huang 2009; Wu, Liu, and Hu 2010; Kim et al 2012) some of which assumed frailty models or proportional hazard models, and used shared random effects between the two processes of longitudinal and survival. However, these models might have convergence issues and could be computationally intensive and difficult to interpret especially when the random effects structures are complex (Wu, Liu, and Hu 2010). Also given that most of these models are built on the assumption of proportionality, when proportionality does not hold, these models lead to biased estimators (Kim, Zeng and Chambless 2012). Other models included the work of Bacci, Bartolucci, and Pandolfi (2016), in which they proposed a joint model for longitudinal and survival data assuming latent process time-varying subject specific random effects that follow first order autoregressive process. However, this model is limited to the assumption of autoregressive processes and violation of this assumption might lead to biased estimates. Li and Su (2017) proposed a likelihood based approach that jointly models the longitudinal outcome and semi-competing events of deaths and dropout, with the assumption that the longitudinal profile conditional on being alive can be expressed in closed form. The dropout time is expressed as a discrete number of follow-up visits prior to dropout. This model is so intensive in the sense that maximization of the likelihood function necessitated calculations of the multivariate normal probabilities of high dimension along with a large number of parameters. In addition, obtaining derivatives of the multivariate normal distribution required numerical computation of the likelihood estimates and Hessian matrix. Most of these approaches discussed thus far did not generate predicted estimates for individual subjects but rather population level estimates.

To avoid the issue of the proportionality assumption of survival models and given that the exact time of dropout is often unavailable, Jaffa et al (2015) proposed an approach where they modelled the number of follow-up visits for every patient prior to dropout as a discrete number instead of assuming the commonly used proportional survival hazard for the censoring process. The number of the last visit for every subject was jointly modeled with the individual's rate of change or slope in one likelihood function. The number of follow-up visits was assumed to be a random variable that follows a discrete distribution, with a mean that is dependent solely on individual slopes without inclusion of the intercepts or accounting for any covariate effect. By ignoring the effect of intercepts, all subjects were treated as if they had the same measurement values at baseline which is not a realistic assumption in clinical settings. In addition, excluding any possible covariate effect on the probability of censoring is not a feasible assumption given that most clinical complications are attributed to a number of factors combined. Moreover, for every subject, the observations were summarized by the individual's ordinary least squared (OLS) estimates for the slope and hence the model required that every subject be seen at least on two occasions. Individuals with only one visit were therefore excluded from the study which introduces bias to the estimates and could result in an underestimation of the population slope. Assuming that the OLS estimates are sufficient statistics for the longitudinal measures and modeling

these estimates instead of the individual observations reduce the computational complexity but might introduce bias to the likelihood function.

In the current study, we relaxed the strong assumptions described earlier and extended the approach of Jaffa et al (2015) by introducing a new likelihood based approach for modeling the longitudinal measures and censoring process. The longitudinal trajectories and baseline measures of the biomarker of interest determined by the individual random slopes and intercepts respectively are both modeled as bivariate random variables with mean dependent on both, the population slopes and intercepts, instead of simply modeling the slopes as univariate normal random effects. The longitudinal profiles for every patient are modeled instead of the OLS estimates to give better insight into disease progression over time and to allow inclusion of all patients together with those with only one measurement.

The dropout time, determined by the number of follow-up visits, is assumed to be a discrete random variable with expected value dependent on selected covariates and on individual random intercepts and slopes of the biomarker of interest. The longitudinal and censoring processes share the individuals' intercepts and slopes which are considered latent random variables in the model. These two processes were jointly modeled in a marginal likelihood function that is integrated over the random effects and maximized to generate maximum likelihood estimates for the population intercept and slope for the outcome of interest and the censoring parameters. Individual slopes were predicted using empirical Bayes estimates.

Our new proposed model has some imbedded computational challenges due to the proliferation in the number of parameters to be estimated and their corresponding high dimensional variance-covariance matrix. Incorporating the individual intercepts along with the slopes requires modeling of a bivariate normal distribution and a variance-covariance matrix instead of a univariate normal distribution with a population estimate for the slope and its variance. Furthermore, modeling longitudinal multiple observations for every subject rather than simplifying these measures to a single OLS estimate, also adds another layer of complexity since it requires maximization of a high dimensional multivariate normal distribution instead of a simple univariate normal distribution. With respect to estimation, incorporating a large number of covariates in the censoring process and joint modeling of the censoring process and the longitudinal measures in one likelihood function make maximization computationally complex and successful convergence a challenge. The performance of this new model and its sensitivity to normality assumptions and underlying distribution of the censoring process were examined using simulation studies and sensitivity analysis. Its feasibility of implementation and successful convergence were also verified. A comparison was also undertaken between this new model and the simpler one of Jaffa et al (2015) to assess if the added level of complexity is superseded by the gain in accuracy of the estimates.

The motivation behind this study is driven by our long standing interest in developing new models to identify biomarkers that predict and associate with progression of renal disease over time. This was achieved by extending the model to incorporate baseline measures through the intercepts, and various covariates in the censoring process and to determine their direct effect on censoring probability. In this regard, we illustrated our new extended model

using a cohort of type 1 diabetic patients that manifest different stages of renal disease. Our interest is centered on identifying the effect of the longitudinal trajectories and baseline measures of plasma prekallikrein (PK) on macroalbuminuria, along with other classical risk factors such as systolic blood pressure (SBP), hemoglobin A1c (HbA1c), and lipid profiles that include total cholesterol, high density lipoprotein (HDL), and low density lipoprotein (LDL), in addition to treatment effect of intensive versus conventional glycemic control.

The new contributions of this article can be presented as such. From a clinical point of view, our study is the first to assess the effect of biomarkers such as PK on the risk of macroalbuminuria-related dropout in type 1 diabetic patients. Moreover, modeling the longitudinal trajectories for every patient gives better insight into progression of renal disease. Our model also overcomes the assumption of proportionality of the hazard, and determines the dropout time by the number of follow-up visits that is dependent on certain specific covariates. The individual intercepts and slopes are both predicted and our likelihood based approach is computationally more feasible compared to other existing approaches.

2. The Model

We propose here an approach for joint modeling of longitudinal measures and censoring process. Interest is in assessing if longitudinal trajectories of a biomarker of interest along with other covariates have any effect on the censoring process. The longitudinal trajectories are determined by individual random intercepts α_i and slopes β_i and censoring process is determined by the number of follow-up visits m_i before reaching the censoring event. The censoring process is a random variable that follows a discrete distribution specified conditionally on the bivariate normally distributed latent random variables (intercepts and slopes) that are shared with the repeated measurements. Moreover, the expected value of the censoring mechanism is assumed to be also dependent on a set of covariates, the effect of which on the probability of censoring is determined here. The model can be specified as follows:

Consider a longitudinal outcome y_{ik} for subject i measured at the k^{th} time point with $i=1, \dots, n$ independent subjects. The measurements were taken at time points $(t_{i1}, t_{i2}, \dots, t_{im_i})$ that do not need to be the same for all patients. Every individual subject i is assumed to have m_i number of follow-up visits before dropping out and m_i is assumed to be $1 \leq m_i \leq p$ where p is denoting the end of study period. In specific, m_i is defined as the number of follow-up visits which reflects the length of stay for subject i in the study before either dropping out or reaching termination of the study. In case of randomly missed intermittent visits, m_i is defined as the number of observed visits plus the visits that are random intermittent missing ($m_i = \text{total number of observed visits} + \text{total number of random intermittent missing visits}$). If a subject does not have any intermittent missing visits then m_i will be equal to the total number of observed visits before dropping out or reaching the end of the study. p is the total number of visits a subject can have before the end of the study. Therefore, if a subject drops out before the end of the study then m_i will be less than p (i.e. $m_i < p$) otherwise the subject will have a complete set of observations and m_i will be equal to p (i.e. $m_i = p$).

The data is assumed to follow a monotone missing pattern; that is subjects are observed until they drop out from the study. Once they drop out they can no longer be followed and hence no further measurements will be recorded on these patients for the duration of the study. The underlying model for y_{ik} is assumed to be

$$y_{ik} = \alpha_i + \beta_i t_{ik} + e_{ik} \quad (1)$$

with random effects α_i and β_i and random error e_{ik} , which is assumed to be normal with mean zero and variance σ_e^2 . Given α_i and β_i y_{ik} is normal with mean $\alpha_i + \beta_i t_{ik}$ and variance σ_e^2 .

Before we present our new model detailed in section 2.1 we first provide a brief overview of the joint modelling approach of Jaffa et al. (2015) that can be described as such: for each individual subject i the specifications in equations (2) to (5) were assumed:

$$a) \beta_i \sim \text{Normal}(\beta, \sigma_\beta^2). \quad (2)$$

$$b) m_i | \beta_i \sim \text{Truncated Discrete Distribution with } \Pr(M = m_i) = F(\gamma_0 + \gamma_1 \beta_i). \quad (3)$$

$$c) b_{i,ols} | m_i, \beta_i \sim \text{Normal}(\beta_i, \sigma_{b_i}^2). \quad (4)$$

In assumption a) the repeated measurements were summarized by only the individual slopes β_i and discarding intercepts. The number of last visit m_i was modeled in b) for subject i and was assumed to be greater than or equal to 2 for all subjects in the study ($2 \leq m_i \leq p$) so that individual slopes $b_{i,ols}$ specified in c) can be estimated. Conditional on the random effect β_i , the distribution for the random variable m_i was assumed to be dependent on the individual slopes without the intercepts and no other covariates were included for computational feasibility. Given that interest was in estimating β_i and knowing that $b_{i,OLS}$ is the ordinary least squares estimate and sufficient statistic for the slope β_i for subject i , the repeated measures y_{ik} for every individual subject i were not modeled but rather were reduced to $b_{i,OLS}$ that were modeled in the likelihood function presented in equation 5. The ordinary least squares estimates $b_{i,ols}$ are assumed in c) to have a normal distribution with mean β_i and variance $\sigma_{b_i}^2$. Hence, the dropout process m_i shared a latent random variable β_i with the repeated measures summarized by $b_{i,ols}$. The corresponding likelihood function for this model (Jaffa et al., 2015) can be described as follows:

$$L = L(\beta, \gamma_0, \gamma_1) = \prod_{i=1}^n \int f(m_i, b_{i,ols}) f(\beta_i) d\beta_i = \prod_{i=1}^n \int f(\beta_i) f(m_i | \beta_i) f(b_{i,ols} | \beta_i, m_i) d\beta_i$$

(5)

This model (Jaffa et al., 2015) imposed two strong assumptions: 1) dropping the intercept and all other covariates from the censoring process and assuming that β_i is the only parameter that affects the censoring process, 2) excluding subjects with only one visit and requiring that every individual has at least two repeated measurements in order to reduce the individual's observations into OLS estimates rather than modeling the longitudinal measures themselves. Our new advanced model that we present in this manuscript (section 2.1) extends the approach of Jaffa et al (2015) and overcomes these limiting assumptions as described in the following section.

2.1 New Extended Model

We present here our new joint modelling approach in which for every individual subject we assume that the longitudinal trajectories of the biomarker of interest determined by individual random intercepts α_i and slopes β_i are modeled as

$$\underline{\beta}_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sim BVN \left\{ \underline{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \Sigma_{\beta} = \begin{bmatrix} \sigma_{\alpha}^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_{\beta}^2 \end{bmatrix} \right\}. \quad (6)$$

Thus, the underlying individual subjects' intercept and slope $\underline{\beta}_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}$ are assumed to follow the bivariate normal distribution with population mean $\underline{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ and variance-covariance

$$\text{matrix } \Sigma_{\beta} = \begin{bmatrix} \sigma_{\alpha}^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_{\beta}^2 \end{bmatrix}.$$

The censoring process is determined by the number of the last visit m_i prior to dropout and is assumed to be dependent on the trajectories of the biomarkers of interest in addition to a set of covariates denoted as X_i . Hence the censoring mechanism can be described as

$$m_i | \underline{\beta}_i \sim \text{Truncated Discrete Distribution with} \quad (7)$$

$$\Pr(M = m_i) = F(\gamma_0 + \underline{\gamma}' \underline{\beta}_i + \underline{\gamma}_3' \underline{X}_i) = F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \underline{\gamma}_3' \underline{X}_i)$$

$$\text{with } \underline{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$$

and $\underline{\gamma}_3$ is a vector of censoring parameters corresponding to the vector of covariates $\underline{X}_i$. Here we specify that conditional on the individual random intercepts and slopes α_i , β_i , the number of follow-up visits m_i is assumed to follow a discrete distribution (such as geometric, and Poisson) with probability dependent on the individual's intercept and slope, in addition to a set of covariates $\underline{X}_i$. The censoring parameters are denoted as γ_0 , γ_1 , γ_2 , and $\underline{\gamma}_3$. Censoring is considered informative with respect to the longitudinal trajectories of a biomarker of interest when either $\gamma_2 \neq 0$ or $\gamma_1 \neq 0$, or both are nonzero. However, censoring is non-informative when these parameters are both zero. If the γ parameter corresponding to a specific covariate X is significant then this covariate has an effect on the censoring process otherwise it does not. The distribution of m_i is right-truncated since the study may terminate before observing the withdrawal of every subject and left truncated since the patient needs to have at least one measurement.

The repeated measures for each individual i are modeled as follows:

$$\underline{Y}_i | m_i, \underline{\beta}_i \sim \text{MVN}(T_i \underline{\beta}_i, \sigma_e^2 I_{m_i}). \quad (8)$$

and the design matrix can be described as such $T_i = \begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{im_i} \end{pmatrix}_{m_i \times 2}$. In assumption (8) we

specify that, given the vector of random effects $\underline{\beta}_i$ and the number of follow-up visits m_i for every subject, the observations for each individual subject $\underline{Y}_i$ follow a multivariate normal distribution with mean $T_i \underline{\beta}_i$ and variance $\sigma_e^2 I_{m_i}$ (Fitzmaurice et al 2009) that is estimated using

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i - \hat{\beta}_i(t_{ij} - \bar{t}_i))^2}{\sum_{i=1}^n m_i - 2n}. \quad (9)$$

2.1.1 Likelihood function and estimation—The likelihood function corresponding to our new extended model described in section 2.1 is described as follows:

$$L = \prod_{i=1}^n f(\underline{\beta}_i, m_i, Y_i) = \prod_{i=1}^n f(Y_i | m_i, \underline{\beta}_i) * f(m_i, \underline{\beta}_i) \Rightarrow \quad (10)$$

$$L = \prod_{i=1}^n f(\underline{\beta}_i) f(m_i | \underline{\beta}_i) f(Y_i | m_i, \underline{\beta}_i).$$

Given that $\underline{\beta}_i$ is not observed so it is considered as the vector of random effects and the likelihood function is integrated over it to get a marginal likelihood that has the following form

$$L = L(\underline{\beta}, \gamma_0, \underline{\gamma}, \Sigma_\beta) = \prod_{i=1}^n \int_{\underline{\beta}_i} f(m_i, Y_i) f(\underline{\beta}_i) d\underline{\beta}_i = \prod_{i=1}^n \int_{\underline{\beta}_i} f(\underline{\beta}_i) f(m_i | \underline{\beta}_i) f(Y_i | \underline{\beta}_i, m_i) d\underline{\beta}_i \quad (11)$$

A direct maximization of the marginal likelihood generates the maximum likelihood estimate (MLE) of the population intercept and slope $\underline{\beta}$ and the censoring parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$. Moreover, empirical Bayes estimates for the individual intercepts and slopes $\underline{\beta}_i$ pertaining to the biomarker of interest are predicted using the adaptive Gaussian quadrature described in Pinheiro and Bates (1995). Direct maximization of the marginal likelihood function was conducted following dual Quasi-Newton optimization algorithm (Fletcher 1987) to obtain consistent estimates of the parameters $(\underline{\beta}, \gamma_0, \underline{\gamma}, \gamma_3, \Sigma_\beta)$. Modeling the individual repeated observations y_{ik} instead of reducing these observations to OLS estimate $b_{i,ols}$ for every individual allows all subjects to be included in the study without exclusion of those with only one visit. Hence our new extended model improves the accuracy of the estimates and eliminates the bias that is introduced by such exclusion and possibly by the OLS estimates themselves. However, this extension adds another level of complexity in the computation along with that attributed to the inclusion of the intercept, its variance and covariance with the slope, in addition to all the covariates $\underline{X}_i$ incorporated in the censoring process. The number of parameters to be estimated has proliferated and the estimation increased from a single parameter σ_β^2 to a matrix of variance-covariance Σ_β , and modeling a multivariate normal distribution for the repeated measures for every individual rather than a simple univariate normal distribution. In this manuscript, two distributions, Geometric and Poisson described in Chattopadhyay et al. (2014) and Plackett (1953), were considered to model the number of visits m_i ; however, other potential distributions can be also followed for the censoring process.

2.1.2 Geometric Distribution Model—Here the number of the last visit for every individual subject is modeled and assumed to follow a truncated geometric distribution with success being determined as subject's dropout of the study. The censoring process can then be defined as follows:

$$m_i | \underline{\beta}_i \sim \text{Truncated geometric with } P_i = F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3' X_i) \quad (12)$$

with probability model

$$\Pr(M = m_i) = \left\{ 1 - F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3' X_i) \right\}^{m_i - 1} \left\{ F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3' X_i) \right\}^{R_i}, m_i = 1, 2, \dots, p$$

(13)

with $F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3' X_i)$ representing the function of individual intercepts and slopes $\underline{\beta}_i$ and the set of covariates X_i . R_i is an indicator variable that was introduced to the geometric distribution to indicate if subject i experienced dropout ($R_i=1$) or remained in the study ($R_i=0$) until its termination time p . Unlike the conventional geometric distribution where success is supposed to occur, the modified geometric distribution presented in this model does not require this assumption since a subject might not dropout from the study. Logistic model may be employed for $F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3' X_i)$ and

$$F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3' X_i) = \frac{1}{1 + \exp \left[-(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3' X_i) \right]}. \quad (14)$$

The likelihood function for subjects with more than one visit can be described as such:

$$\begin{aligned}
L &= L(\underline{\beta}, \gamma_0, \underline{\gamma}, \underline{\gamma}_3, \Sigma_\beta) = \prod_{i=1}^n \int_{\underline{\beta}_i} f(m_i, Y_i) f(\underline{\beta}_i) d\underline{\beta}_i = \prod_{i=1}^n \int_{\underline{\beta}_i} f(\underline{\beta}_i) f(m_i | \underline{\beta}_i) f(Y_i | \underline{\beta}_i, m_i) d\underline{\beta}_i \\
&= \prod_{i=1}^n \int_{\underline{\beta}_i} \exp \left(-\frac{1}{2} \left[\frac{1}{1 - \rho_{\alpha_i, \beta_i}^2} \left(\frac{\alpha_i - \alpha}{\sigma_\alpha} \right)^2 - 2\rho_{\alpha_i, \beta_i} \left(\frac{\alpha_i - \alpha}{\sigma_\alpha} \right) \left(\frac{\beta_i - \beta}{\sigma_\beta} \right) + \left(\frac{\beta_i - \beta}{\sigma_\beta} \right)^2 \right] \right) \\
&\quad * \left\{ 1 - F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \underline{\gamma}_3' X_i) \right\}^{m_i - 1} * \left\{ F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \underline{\gamma}_3' X_i) \right\}^{R_i} \\
&\quad * \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp \left\{ \frac{-1}{2} \left(\frac{Y_i - T_i \beta_i}{\sigma_\varepsilon} \right)^2 \right\} d\underline{\beta}_i
\end{aligned} \tag{15}$$

The log-likelihood can be expressed as

$$\begin{aligned}
\ell \alpha - \int_{\underline{\beta}_i} \frac{1}{2} \sum_{i=1}^n (\underline{\beta}_i - \beta)' \Sigma_\beta^{-1} (\underline{\beta}_i - \beta) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n (Y_i - T_i \beta_i)' (Y_i - T_i \beta_i) \\
+ \sum_{i=1}^n (m_i - 1) \log \left(1 - F(\gamma_0 + \gamma' \beta_i + \underline{\gamma}_3' X_i) \right) + \sum_{i=1}^n R_i \log \left(1 - F(\gamma_0 + \gamma' \beta_i + \underline{\gamma}_3' X_i) \right) d\underline{\beta}_i
\end{aligned} \tag{16}$$

Since subjects with only one observation have slopes of zero, their contribution to the likelihood function in equation (15) would be through their baseline measure reflected by the intercept (Glidden and Wei 1999). Specifically, their contribution to the likelihood function is described as such:

$$\begin{aligned}
L &= L(\underline{\beta}, \gamma_0, \underline{\gamma}, \Sigma_\beta) = \prod_{i=1}^n \int_{\underline{\beta}_i} f(m_i, Y_i) f(\underline{\beta}_i) d\underline{\beta}_i = \prod_{i=1}^n \int_{\underline{\beta}_i} f(\underline{\beta}_i) f(m_i | \underline{\beta}_i) f(Y_i | \underline{\beta}_i, m_i) d\underline{\beta}_i \tag{17} \\
&= \prod_{i=1}^n \int_{\alpha_i} \exp \left(-\frac{1}{2} \left[\left(\frac{\alpha_i - \alpha}{\sigma_\alpha} \right)^2 + \left(\frac{\beta}{\sigma_\beta} \right)^2 \right] \right) * \left\{ F(\gamma_0 + \gamma_1 \alpha_i + \underline{\gamma}_3' X_i) \right\} * \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \\
&\quad \exp \left\{ \frac{-1}{2} \left(\frac{Y_i - \alpha_i}{\sigma_\varepsilon} \right)^2 \right\} d\alpha_i
\end{aligned}$$

The log of the marginal likelihood function (eq. 15, and 16) is maximized as previously described and maximum likelihood estimates of the population intercept and slope (α and

β), censoring parameters $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ and variance-covariance matrix Σ_β are all generated and empirical Bayes estimates for individual intercepts and slopes (α_i and β_i) for each subject are predicted.

2.1.3 Poisson Distribution Model—Here the censoring process is assumed to follow Poisson distribution with mean dependent on the individual intercepts, slopes and covariates $\underline{X}_i$. The Poisson distribution is left and right truncated as was previously specified and has the following probability model:

$$\Pr [M = m_i] = \frac{\mu_i^{(m_i-1)} e^{-\mu_i}}{(m_i-1)! * \Pr (1 \leq m_i \leq p)} \quad (18)$$

with $m_i = 1, 2, \dots, p$ and mean $\mu_i = \exp(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i + \gamma_3 \underline{X}_i)$. The likelihood function corresponding to the Poisson model for individuals with more than one visit is described as follows:

$$\begin{aligned} L = L(\underline{\beta}, \gamma_0, \underline{\gamma}, \Sigma_\beta) &= \prod_{i=1}^n \int_{\underline{\beta}_i} f(m_i, \underline{Y}_i) f(\underline{\beta}_i) d\underline{\beta}_i = \prod_{i=1}^n \int_{\underline{\beta}_i} f(\underline{\beta}_i) f(m_i | \underline{\beta}_i) f(\underline{Y}_i | \underline{\beta}_i, m_i) d\underline{\beta}_i \quad (19) \\ &= \prod_{i=1}^n \int_{\underline{\beta}_i} \exp \left(-\frac{1}{2} \left[\frac{1}{1 - \rho_{\alpha_i, \beta_i}^2} \left\{ \left(\frac{\alpha_i - \alpha}{\sigma_\alpha} \right)^2 - 2\rho_{\alpha_i, \beta_i} \left(\frac{\alpha_i - \alpha}{\sigma_\alpha} \right) \left(\frac{\beta_i - \beta}{\sigma_\beta} \right) + \left(\frac{\beta_i - \beta}{\sigma_\beta} \right)^2 \right\} \right] \right) \\ &\quad * \frac{\mu_i^{(m_i-1)} e^{-\mu_i}}{(m_i-1)! * \Pr (1 \leq m_i \leq p)} \\ &\quad * \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp \left\{ -\frac{1}{2} \left(\frac{\underline{Y}_i - T_i \underline{\beta}_i}{\sigma_\epsilon} \right)^2 \right\} d\underline{\beta}_i \end{aligned}$$

The log-likelihood is expressed as

$$\begin{aligned} \ell \alpha \int_{\underline{\beta}_i} & -\frac{1}{2} \sum_{i=1}^n (\underline{\beta}_i - \beta)' \Sigma_\beta^{-1} (\underline{\beta}_i - \beta) - \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^n (\underline{Y}_i - T_i \underline{\beta}_i)' (\underline{Y}_i - T_i \underline{\beta}_i) \quad (20) \\ & + \sum_{i=1}^n (m_i - 1) \log(\mu_i) - \mu_i d\underline{\beta}_i \end{aligned}$$

The contribution of individuals with only one observation to the likelihood function in equation (19) is defined as follows:

$$\begin{aligned}
L = L(\underline{\beta}, \gamma_0, \underline{\gamma}, \Sigma_\beta) &= \prod_{i=1}^n \int_{\underline{\beta}_i} f(m_i, Y_i) f(\underline{\beta}_i) d\underline{\beta}_i = \prod_{i=1}^n \int_{\underline{\beta}_i} f(\underline{\beta}_i) f(m_i | \underline{\beta}_i) f(Y_i | \underline{\beta}_i, m_i) d\underline{\beta}_i \quad (21) \\
&= \prod_{i=1}^n \int_{\underline{\beta}_i} \exp \left(-\frac{1}{2} \left[\left(\frac{\alpha_i - \alpha}{\sigma_\alpha} \right)^2 + \left(\frac{\beta}{\sigma_\beta} \right)^2 \right] \right) * \frac{e^{-\mu_i}}{\Pr(1 \leq m_i \leq p)} \\
&\quad * \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp \left\{ -\frac{1}{2} \left(\frac{Y_i - \alpha_i}{\sigma_\epsilon} \right)^2 \right\} d\alpha_i
\end{aligned}$$

with $\mu_i = \exp(\gamma_0 + \gamma_1 \alpha_i + \gamma_3 X_i)$. Maximization of the log of the marginal likelihood function (eq. 19 and 20) results in generating estimates for the parameters of interest specified earlier in the geometric distribution section.

3. Type 1 Diabetes Application

The new extended model was applied on a cohort of 569 type 1 diabetic patients enrolled in the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) study (1986 (1999)). One of the major complications in this cohort of patients is diabetic nephropathy characterized by sustained macroalbuminuria. The levels of AER were used as a risk indicator to monitor progression of renal disease in diabetic patients. Ranges for AER levels can be categorized as such: 0–39 mg/24h (normoalbuminuria); 40–300 mg/24h (microalbuminuria) a biomarker for early renal disease and AER > 300 mg/24h (macroalbuminuria), a biomarker for advanced stage kidney disease. Here, patients were considered to be informatively right censored when their AER levels exceeded 300 mg/hr at first onset. Five repeated measurements for AER were taken between years 1983 and 2005 and were indexed as follow-up years 0, 10, 16, 20, and 22. Diabetic patients who did not develop macroalbuminuria had a complete set of five AER measurements and other clinical parameters; whereas, patients with macroalbuminuria had informative right censoring and an incomplete set of AER measures and clinical parameters. A total of 101 patients (18%) developed macroalbuminuria and were therefore right censored. The main objective here is to assess the differential effect of longitudinal trajectories of plasma PK as well as its baseline values on macroalbuminuria. Hence, its corresponding slopes and intercepts were included as covariates in the censoring process along with SBP, HbA1c, the lipid profiles and treatment. The longitudinal trajectories of plasma PK were modeled in equation (21) as a function of time used on a logarithmic scale to achieve linearity.

$$\log(Plasma PK_{ik}) = \alpha_{i_Plasma_PK} + \beta_{i_Plasma_PK} \log(\log(year)_{ik} + 1) + e_{ik} \quad (22)$$

The latent parameters; intercepts $\alpha_{i_Plasma_PK}$ and slopes $\beta_{i_Plasma_PK}$, are shared with the censoring process that has also other factors as defined in equation (23)

$$\Pr(M = m_i) = F(\gamma_0 + \gamma'_1 \beta_i + \gamma'_3 X_i) = F(\gamma_0 + \gamma_1 \alpha_{i_Plasma_PK} + \gamma_2 \beta_{i_Plasma_PK} + \gamma_3 SBP_i + \gamma_4 HbA1c + \gamma_5 HDL + \gamma_6 LDL + \gamma_7 Total_Cholesterol + \gamma_8 Treatment) \quad (23)$$

Significance of γ_1 and γ_2 indicates interdependency between the longitudinal trajectories of plasma PK and the macroalbuminuria censoring process. If all of the γ 's are non-significant then this indicates that the censoring is non-informative. However, if any of the γ 's are significant, then the dropout is considered informative and dependent on the factors pertaining to the significant γ 's.

Selection of models for the censoring process was based on the measures of fit for the dataset using Akaike information criterion (AIC), and Bayesian information criterion (BIC). The Poisson censoring model had lower AIC = 3454, and BIC = 3399 compared to the geometric censoring model that had AIC = 7427, and BIC = 7487. This indicates that Poisson is a better fit for this dataset than the geometric model. The results generated under the Poisson model showed that plasma PK was increasing significantly over time with $\hat{\beta} = 0.086$ (SE=0.005, P-value<0.0001) and intercept $\hat{\alpha} = 4.156$ (SE=0.013, P-value<0.0001). The longitudinal trajectories of plasma PK denoted as individual slopes $\beta_{i_Plasma_PK}$ were found to be significantly associated with the censoring process ($\hat{\gamma}_2 = 0.801$, SE = 0.073, P-value = 0.0321) adjusting for the effect of SBP, HbA1c, lipid levels (HDL, LDL, total cholesterol) and treatment. This result indicates that the higher the changes in plasma PK the longer the dropout time. However, the baseline measures denoted as $\alpha_{i_Plasma_PK}$ were not found to be significantly associated with the censoring process (P-value = 0.2778). SBP ($\hat{\gamma}_3 = -0.005$, SE = 0.0021, P-value = 0.0131) and HbA1c ($\hat{\gamma}_4 = -0.078$, SE = 0.0167, P-value < 0.0001) were both significantly associated with macroalbuminuria, wherein increased levels in their values was shown to be associated with a decrease in the number of follow-up visit m_i and faster onset of macroalbuminuria. Lipids cholesterol (total-cholesterol, HDL, and LDL) and treatment were not associated with the dropout events. Hence, the censoring process macroalbuminuria was shown to be informative and dependent on the longitudinal trajectories of plasma PK, and is also a function of SBP and HbA1c. Similar results were obtained under the geometric model that showed that plasma PK was increasing significantly over time with $\hat{\beta} = 0.093$ (SE=0.005, P-value<0.0001) and intercept $\hat{\alpha} = 4.155$ (SE=0.012, P-value<0.0001). The baseline measures and longitudinal trajectories of plasma PK were both found to be significantly associated with the censoring process (P-value < 0.0001) adjusting for the effect of SBP, HbA1c, lipid levels (HDL, LDL, total cholesterol) and treatment. In this model, only HbA1c was found to be a significant predictor of the censoring probability of macroalbuminuria. However, SBP, lipid profiles and treatment were not associated with the censoring process after adjusting for the trajectories of plasma PK, and HbA1c. Figures 1 and 2 depict the predicted empirical Bayes estimates for individual intercepts and slopes as well as their respective means pertaining to log plasma PK levels plotted against the number of visits m_i adjusting for the effect of the clinical covariates. Figure 1 shows that patients with shorter length of stay ranging between 1 and 2 visits had higher estimates for the intercepts and therefore increased baseline values of plasma PK compared to those with longer time to macroalbuminuria ranging between 3 and 4 visits.

Similarly, in Figure 2 patients who experienced macroalbuminuria had increased values of the slope estimates for plasma PK longitudinal trajectories compared to those who did not experience this event. Thus, one can deduce that longitudinal trajectories of plasma PK are independent predictors to developing macroalbuminuria, after adjusting for known classical clinical risk factors such as HbA1c, blood pressure, lipid profiles and treatment.

4. Simulation Study

A simulation study was conducted to assess the performance of the model for both distributions, Poisson and geometric. The censoring process was defined as follows

$$\Pr(M = m_i) = F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i) = F(\gamma_0 + \gamma_1 \alpha_i + \gamma_2 \beta_i). \quad (24)$$

The function F is determined as geometric or Poisson model as was described earlier in section 2. Different levels of censoring were considered and varied from low level with $\gamma_2 = -1.4$, mid level with $\gamma_2 = -5.2$ and high level with $\gamma_2 = -10.5$ (Tables 1 and 3, Column 3). Sensitivity analysis was conducted to assess robustness of the model when censoring is non-informative with $\gamma_2 = 0$ and $\gamma_1 = 0$ or dependent on either the individual intercepts or slopes but not on both (Tables 2 and 4), and when underlying distribution of the censoring process was misspecified (Tables 1 to 4 Column 4). A comparison was also conducted with the model of Jaffa et al. (2015) described in section 2. The correct underlying distribution for the censoring process was assumed for this model and its performance indicators are presented in column 5 of Tables 1 to 4. The objective of this comparison is to assess if the added level of complexity imbedded in our new model proposed in section 2.1 is compensated by increased level of accuracy. Performance was assessed using bias, MSE(a) mean squared errors for the population intercepts and slopes (α , and β), and MSE(b) mean squared errors for the individual intercepts and slopes (α_i , and β_i). These performance indicators for intercepts and slopes can be described as such:

$$MSE(a)_{-\alpha} = \frac{1}{r} \sum_{s=1}^r (\hat{\alpha}_s - \alpha)^2 \quad (25)$$

$$MSE(a)_{-\beta} = \frac{1}{r} \sum_{s=1}^r (\hat{\beta}_s - \beta)^2 \quad (26)$$

$$MSE(b)_{-\alpha} = \frac{1}{nr} \sum_{s=1}^r \sum_{i=1}^n (\hat{\alpha}_{is} - \alpha_{is})^2 \quad (27)$$

$$MSE(b)_{-\beta} = \frac{1}{nr} \sum_{s=1}^r \sum_{i=1}^n (\hat{\beta}_{is} - \beta_{is})^2 \quad (28)$$

Intercepts and slopes $\underline{\beta}_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}$ were simulated from a bivariate normal distribution with a covariance of $\sigma_{\alpha, \beta} = -0.02$ and the individual observations were generated from the following equation

$$y_{ik} = \alpha_i + \beta_i \log(t_{ik} + 1) + e_{ik} \quad (29)$$

where e_{ik} is a random error assumed to be normal with mean zero and variance $\sigma_e^2 = 0.12$. The detailed values of the parameters used in our simulation study are presented in the footnote for Tables 1 to 4. We generated 2000 Monte Carlo simulated datasets each with a sample size of 200 subjects. Our simulation results suggested that both distributions have overall similar performance in terms of bias and MSEs under the various levels of censoring. The bias associated with the estimates generated under the Poisson model ranged from 0.0023 to 0.0104 in absolute terms, MSE(a) ranged from 0.0079 to 0.0195, and MSE(b) from 0.0021 to 0.0198 (Tables 1 and 2, Column 3). The performance indicators for the geometric model ranged between 0.0033 to 0.0084 for bias, 0.0098 to 0.0173 for MSE(a) and 0.0036 to 0.0099 for MSE(b) (Tables 3 and 4, Column 3). Hence geometric model appeared to have relatively slightly better performance indicators than the Poisson model. Our sensitivity analysis suggested that when censoring is dependent only on intercepts ($\gamma_2 = 0$, $\gamma_1 \neq 0$) the performance of the model under both distributions still resulted in small performance indicators in comparison to the correct assumption of informative censoring where both $\gamma_2 \neq 0$, $\gamma_1 \neq 0$ (see Tables 2 and 4, Column 3, where $\gamma_2 = 0$ and $\gamma_1 \neq 0$). However, compared to geometric, Poisson model had larger bias and MSEs by about 2 fold under this particular censoring in the majority of the estimates. Similarly, when the censoring mechanism was dependent only on individual slopes ($\gamma_2 \neq 0$, $\gamma_1 = 0$) and not on intercepts, both distributions had small performance indicators that are very comparable to the correct case of informative censoring that is dependent on both parameters. Poisson distribution continued to have higher bias and MSEs in comparison to geometric distribution by about 2 folds suggesting again that it could be more sensitive to the violation of censoring assumptions. When the censoring process was assumed to be non-informative ($\gamma_2 = 0$, $\gamma_1 = 0$) both distributions showed robustness to the violation of underlying assumption about the censoring mechanism (Tables 2 and 4). In this regard, both distributions continued to have small bias and MSEs that are comparable to those generated under the correct assumptions, and Poisson had higher values for the performance indicators compared to Geometric. Hence, Geometric distribution appeared to have better precision and accuracy in the estimation compared to Poisson. Nevertheless, irrespective of the assumed distribution the proposed model resulted in accurate estimates under the different levels of censoring.

Robustness of the model for distribution misspecification was also assessed in a sensitivity analysis in which censoring data were simulated from Poisson distribution but geometric model was assumed for the censoring process (Tables 1 and 2, Column 4) and similarly data were simulated from geometric distribution but Poisson model was assumed (Tables 3 and 4, Column 4). Our results suggested that when data were simulated as Poisson but the censoring process was treated as geometric, the associated estimates had an observed increase in bias and MSEs by about 2 fold or less for the majority of the estimates in comparison to the correct assumption of Poisson. Exceptionally, $MSE(b)_{\beta}$ at $\gamma_2 = -5.2$ had an observed increase that is due to the distribution misspecification by about 4.25 fold and that for $\gamma_2 = -10.5$ by about 4 fold (Tables 1 and 2, Column 4).

When data were simulated from a geometric distribution but the Poisson distribution was assumed for the censoring process (Tables 3 and 4, Column 4), an observed increase of about 2 fold or less compared to the correct assumption of geometric was denoted for the majority of the estimates with the exception of $MSE(b)_{\beta}$ at $\gamma_2 = -1.4$ at that exhibited an increase of about 4 fold. As anticipated, it appears that the model is sensitive to the assumption of distribution for the censoring process; however, the performance indicators generated under the misspecification appeared to be still minimal despite the observed increase in its values. Therefore, the model could still be considered robust for the underlying assumptions for the censoring process, but using the correct distribution is definitely more efficient. One can base the choice of the distribution for the censoring process using measures of fit. The distribution that results in the smallest AIC and BIC is typically the one that fits the data best.

We also carried out a comparison between our new extended model described in section 2.1 and that of Jaffa et al. (2015) described in section 2 that assumes dependence of the censoring process on the slopes β_i only, and does not include the intercepts α_i . Hence intercepts are not estimated under this model and their corresponding performance indicators are not generated. Results of the simulation study conducted on this simpler model are presented in column 5 of Tables 1 to 4. Under both distributions our new model showed superiority in performance compared to the simpler model. In specific, bias in the slope estimation has increased by 2 to 3 folds on average under the simpler model with geometric and Poisson distributions respectively compared to the new model. Similarly, attributed $MSE(a)$ and $MSE(b)$ for the slopes also increased by about 2 folds on average under the simpler model for both distributions. Hence, implementing our new model improved the accuracy and precision of the estimates by at least 2 folds on average, and this was demonstrated at all levels of censoring. This gain in accuracy is due to the fact that the likelihood function is using information from both intercepts and slopes, and is accounting for their correlation. In addition, using information from longitudinal measures instead of OLS reduces the bias that is attributed to these estimates and introduced to the likelihood function when OLS were modeled instead of the longitudinal trajectories. Moreover, including observations from subjects with only 1 measure and allowing these subjects to contribute to information in the likelihood function also increases the level of accuracy.

A sensitivity analysis was also conducted to assess the effect of non-normality on the performance of the model. Specifically, we simulated the errors e_{ik} as non-normal and

following a triangular distribution with a mean of 0, mode of 1, minimum of -3 and maximum of 2 and not symmetric about 0. A triangular distribution with these particular values of the mode, mean, maximum and minimum should have a moderate to large departure from normality, and conditional on β_i , the outcome y_{ik} follows a triangular distribution. A medium level of censoring of -5.2 was assumed. Our results showed some increase in the performance indicators as compared to the normal error for both Poisson and geometric distributions. Specifically, for the Poisson distribution, the associated bias for the intercept and slope was -0.0097 and 0.00829 respectively, $MSE(a)$ was 0.03786 and 0.01412 , and $MSE(b)$ was 0.02065 and 0.0038 . For the geometric distribution, the bias for the intercept and slope was respectively -0.00859 and 0.00646 , $MSE(a)$ was 0.01109 and 0.01680 , and $MSE(b)$ was respectively 0.01031 and 0.01271 . Accordingly, one can deduce that for both distributions the impact of violating normality assumption by having triangular errors was very minimal and less than 2 fold increase on the majority of the performance indicators compared to the normally distributed data. Thus, this result suggests robustness of our new model for whether the censoring process was informative or non-informative, and for normality assumptions of the outcome. In addition, our results showed some robustness for the censoring distribution, but in all cases it is more efficient to use the correct dropout model. More importantly, we verified that excluding subjects with only one measure, and not accounting for the intercepts in the model, and not modeling the longitudinal trajectories contribute to a decrease in the accuracy of the estimates.

5. Discussion

Conventional methods used to analyze longitudinal data and ignore informative right censoring or treat it as non-informative are inefficient and result in biased estimates. Available approaches had it be the ones based on joint modeling of time-to-event and longitudinal measurements, Bayesian or semiparametric-based represent methodological advance but have restrictive assumptions that are not well specified and are characterized by inherent complexity of implementation (Tsiatis and Davidian 2001). Likelihood approaches based on specifying a likelihood function for the parameters of the longitudinal process and models for the censoring events were used to avoid approximation. In this regard EM algorithms were developed to maximize the joint likelihood function in which parametric distributions were assumed for the censoring process and the shared random effects (DeGruttola and Tu 1994; Wulfoshn and Tsiatis 1997). Nevertheless, available likelihood approaches require solving equations with intractable integrals for the random effects and infinite-dimensional proportional hazards model, and complexity increases in the presence of missing data (Wu, Liu, and Hu 2010). Approaches such as those by Wu and Carroll (1998) and Wu and Bailey (1998) in which the censoring process was assumed to follow the probit distribution (Wu and Carroll 1998) and slopes were generated using linear minimum mean square error and minimum variance unbiased estimators (Wu and Bailey 1998), had innate assumptions that were not lax and were limited to generating population slopes only and not individual slopes. Even though joint modeling has been commonly used to link longitudinal and survival data, several limitations were encountered that include complexity of calculations that arise from the integration of the random effect (Gould et al, 2015), and convergence issues especially with semiparametric and nonparametric models. Models that

account for missing data such as but not limited to Wu, Hu, and Wu (2008), Liu, Huang, and O'Quigley (2008), Liu and Huang (2009), and Kim et al (2012), considered joint modeling of Linear or nonlinear mixed effects and Cox-proportional hazards. These models were implemented using EM algorithms for the joint inferences, and intensive computations were experienced (Wu, Liu and Hu 2010). More importantly, in case proportionality assumption was violated these models are known to result in biased estimates.

Here we present an advanced likelihood based approach for joint modeling of the longitudinal measures and censoring process. Changes in the longitudinal trajectories of a biomarker of interest and its baseline measures summarized in terms of individual slopes and intercepts are modeled as bivariate random variables and are shared with the censoring process. To overcome the issue of non-proportionality in the hazard function, we assumed that the dropout time, as in Li and Su (2017), to be the number of follow-up visits since the exact time of dropout is not known in most practical situations. We extended the censoring process to include the rate of change in trajectories and baseline measures of the biomarker of interest, and a set of covariates that could have direct effect on the censoring probability. The longitudinal repeated measures for every subject were jointly modeled with the censoring process along with the individual random intercepts and slopes in a likelihood function that is integrated over the random effects. The marginal likelihood function is maximized using dual Quasi-Newton optimization algorithm (Fletcher 1987). Maximum likelihood estimates for the population intercepts and slopes, their variance-covariance matrix, and censoring parameters of the individual intercepts and slopes, and the covariates included in the dropout mechanism are all generated. Empirical Bayes estimates for the individual subject random effects were predicted using adaptive Gaussian quadrature described in Pinheiro and Bates (1995). Given that individual random intercepts and slopes are shared between the censoring process and the longitudinal measures, they are considered latent parameters in this model.

The advantage of this new approach is that it provides unbiased estimates for the population and individual intercepts and slopes to reflect the trajectories over times for the biomarker of interest adjusting for informative right censoring. In addition it gives direct assessment of the predictive utility of the covariates that were considered in the censoring process, the baseline measures and longitudinal trajectories of the biomarker, on the risk of censoring, and overcomes the strong assumption of proportional hazards that is commonly used in the literature. Hence our model presents a tool that can be used to identify novel biomarkers as protective or risk factors for developing a certain clinical end point. In this regard, when this model was illustrated using the DCCT/EDIC-cohort of type 1 diabetic patients, trajectories of plasma PK were modeled and their effects were shown to be directly related to macroalbuminuria. Hence, the predictive utility of plasma PK was demonstrated as a risk factor for macroalbuminuria and its direct effect on the censoring process was shown in the presence of the classical risk factors; blood pressure, HbA1c, lipid levels and treatment. This application presents one aspect of the widespread utility of the model on diverse clinical datasets where interest is centered on identifying biomarkers and factors that are attributing to patients' dropout and probability of onset of a specific clinical event.

The accuracy of the estimates for the individual and population intercepts and slopes for the biomarker of interest was established using extensive simulation study. Our simulation results showed superiority of this model over the simpler one by Jaffa et al (2015) in addition to its robustness to the normality assumptions (example triangular errors). Misspecification of the underlying assumption for the censoring process increased bias and MSE by maximum of 2 folds. Even though these values were still small, using the accurate distribution is favorable in order to ensure accuracy of the results and inferences.

One can conclude that the added level of computational complexity introduced by the proliferation of the number of parameters and their corresponding variance-covariance matrix and modeling of a multivariate normal distribution was outweighed by the gain in accuracy and the practical utility of this new model.

Acknowledgments

This work was supported by the National Institutes of Health Grants HL077192 (AAJ), and 5 P01 HL055782. We would like to acknowledge the DCCT/EDIC as the source of data. A complete list of participants in the DCCT/EDIC Research Group is presented in the Supplementary Material published online for the article in *N Engl J Med* 2015;372:1722–33.

References

- Bacci S, Bartolucci F, Pandolfi S. 2016; A joint model for longitudinal and survival data based on an AR(1) latent process. *Statistical Methods in Medical Research*. doi: 10.1177/0962280216659895
- Brown ER, Ibrahim JG. 2003; A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*. 59:221–228. [PubMed: 12926706]
- Brown ER, Ibrahim JG. 2003; Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics*. 59:686–93. [PubMed: 14601770]
- Chattopadhyay S, Murthy CA, Pal SK. 2014; Fitting truncated geometric distributions in large scale real world networks. *Theoretical Computer Science*. 551:22–38.
- DeGruttola V, Tu XM. 1994; Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*. 50:1003–14. [PubMed: 7786983]
- Ding J, Wang JL. 2008; Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*. 64:546–556. [PubMed: 17888040]
- EDIC Research Group: Epidemiology of diabetes interventions and complications (EDIC). 1999; Design, implementation and preliminary results of long-term follow up of diabetes control and complications trial cohort. *Diabetes Care*. 22:99–111. [PubMed: 10333910]
- Faucett CJ, Thomas DC. 1996; Simultaneously modeling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine*. 15:1663–1685. [PubMed: 8858789]
- FitzmauriceGD, Davidian, M, Verbeke, G, Molenberghs, G. *Longitudinal Data Analysis*. Chapman and Hall/CRC; 2009.
- Fletcher, R. *Practical Methods of Optimization*. 2. New York: Wiley; 1987.
- Gallant AR, Nychka DW. 1987; Semi Nonparametric maximum likelihood estimation. *Econometrica*. 55:363–390.
- Glidden DV, Wei LJ. 1999; Rank estimation of treatment differences based on repeated measurements subject to dependent censoring. *Journal of the American Statistical Association*. 94:888–895.
- Gould LA, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, Bois FY. 2015; Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine*. 34:2181–2195. [PubMed: 24634327]

- Guo X, Carlin BP. 2004; Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*. 58:1–9.
- Henderson R, Diggle P, Dobson A. 2000; Joint modeling of longitudinal measurements and event time data. *Biostatistics*. 4:465–480.
- Hsieh F, Tseng Y, Wang J. 2006; Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*. 62:1037–1043. [PubMed: 17156277]
- Huang X, Liu L. 2007; A joint frailty model for survival and gap times between recurrent events. *Biometrics*. 63:389–397. [PubMed: 17688491]
- Jacqmin-Gadda H, Proust-Lima C, Taylor JMG, Commenges D. 2010; Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics*. 66:11–19. [PubMed: 19432771]
- Jaffa AA, Durazo-Arvizu R, Zheng D, Lackland DT, Srikanth S, Garvey WT, Schmaier AH. the DCCT/EDIC Study Group. 2003; Plasma Prekallikrein: A Marker for hypertension and nephropathy in type 1 diabetes. *Diabetes*. 52:1215–1221. [PubMed: 12716755]
- Jaffa MA, Lipsitz S, Woolson RF. 2015; Slope Estimation for Informatively Right Censored Longitudinal Data Modeling the Number of Observations Using Geometric and Poisson Distributions: Application to Renal Transplant Cohort. *Statist Meth Med Res*. 24:819–835.
- Jaffa MA, Luttrell D, Schmaier AH, Klein RL, Lopes-Virella M, Luttrell LM, Jaffa AA. DCCT/EDIC Research Group. 2016; Plasma Prekallikrein is Associated with Carotid Intima-Media Thickness in Type 1 Diabetes. *Diabetes*. 65:498–502. [PubMed: 26603531]
- Kim S, Zeng D, Chambless L, Li Y. 2012; Joint models of longitudinal data and recurrent events with informative terminal event. *Statistics in Bioscience*. 4:262–281.
- Li Q, Su L. 2017 Accommodating informative dropout and death: a joint modeling approach for longitudinal and semicompeting risks data. *Journal of the Royal Statistical Society, Series C*.
- Lin DY, Ying Z. 2001; Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*. 96:103–113.
- Liu L, Huang X, O'Quigley J. 2008; Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*. 64:950–958. [PubMed: 18162110]
- Liu L, Huang X. 2009; Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society, Series C*. 58:65–81.
- Liu L, Wolfe R, Huang X. 2004; Shared frailty models for recurrent events and a terminal event. *Biometrics*. 60:747–756. [PubMed: 15339298]
- Molitch ME, Steffes M, Sun W, Rutledge B, Cleary P, de Boer IH, Zinman B, Lachin J. 2010; Epidemiology of Diabetes Interventions and Complications Study Group. Development and progression of renal insufficiency with and without albuminuria in adults with type 1 diabetes in the diabetes control and complications trial and the epidemiology of diabetes interventions and complications study. *Diabetes Care*. 33:1536–43. [PubMed: 20413518]
- Nathoo FS, Dean CB. 2008; Spatial multistate transitional models for longitudinal event data. *Biometrics*. 64:271–279. [PubMed: 17425639]
- Pan Q, Yi GY. 2011; An estimation method of marginal treatment effects on correlated longitudinal and survival outcomes. *Statistics and Its Inference*. 4(4):499–509.
- Pinheiro JC, Bates DM. 1995; Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. *Journal of Computational and Graphical Statistics*. 4:12–35.
- Plackett RL. 1953; The Truncated Poisson Distribution. *Biometrics*. 9(4):485–488.
- Rizopoulos D, Verbeke G, Lesaffre E. 2009; Fully exponential Laplace approximations for the joint modeling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B*. 71:637–654.
- Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. 2007; Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*. 8:708–721. [PubMed: 17267392]
- Song X, Davidian M, Tsiatis AA. 2002; A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*. 58:742–753. [PubMed: 12495128]

- Song X, Wang CY. 2008; Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics*. 64:557–566. [PubMed: 17725812]
- Taylor JMG, Cumberland WG, Sy JP. 1994; A stochastic model for analysis of longitudinal data. *Journal of the American Statistical Association*. 89:727–76.
- The DCCT Research Group. The diabetes control and complications trial (DCCT). 1986; Design and methodologic considerations for the feasibility phase. *Diabetes*. 35:530–545. [PubMed: 2869996]
- Tsiatis AA, Davidian MA. 2001; Semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*. 88:447–58.
- Tsiatis AA, Davidian MA. 2004; An overview of joint modeling of longitudinal and time-to-event data. *Statistica Sinica*. 14:793–818.
- Tsiatis AA, DeGruttola V, Wulfsohn MS. 1995; Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*. 90:27–37.
- Vonesh E, Greene T, Schluchter M. 2006; Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*. 25:143–163. [PubMed: 16025541]
- Wang M, Qin J, Chiang C. 2001; Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*. 96:1057–1065.
- Wang Y, Taylor JMG. 2001; Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*. 96:895–905.
- Wu L, Hu XJ, Wu H. 2008; Joint inference for nonlinear mixed-effects models and time-to-event at the presence of missing data. *Biostatistics*. 9:308–320. [PubMed: 17728318]
- Wu L, Liu W, Hu XJ. 2010; Joint inference on HIV viral dynamics and immune suppression in presence of measurement errors. *Biometrics*. 66:327–335. [PubMed: 19673859]
- Wu MC, Bailey K. 1988; Analyzing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*. 7:337–346. [PubMed: 3281208]
- Wu MC, Carroll RJ. 1998; Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 44:175–188.
- Wulfsohn MS, Tsiatis AA. 1997; A joint model for survival and longitudinal data measured with error. *Biometrics*. 53:330–339. [PubMed: 9147598]
- Xu J, Zeger SL. 2001; Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*. 50:375–387.
- Yao Q, Wei LJ, Hogan JW. 1988; Analysis of incomplete repeated measurements with dependent censoring times. *Biometrika*. 85:139–149.
- Ye Y, Kalbfleisch J, Schaubel D. 2007; Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*. 63:78–87. [PubMed: 17447932]
- Zhang D, Lin X, Raz J, Sowers M. 1998; Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*. 93:710–719.

Figure 1.

Individual and mean values of predicted empirical Bayes estimates for individual intercepts of logarithmic plasma prekallikrein plotted against the number of last visits adjusted for SBP, Lipids and treatment effect. Mean values are represented by the diamond symbol.

Figure 2.

Individual and mean values of predicted empirical Bayes estimates for individual slopes of logarithmic plasma prekallikrein plotted against the number of last visits adjusted for SBP, Lipids and treatment effect. Mean values are represented by the diamond symbol.

Table 1

Summary * of Performance of Three Estimators: 1) Likelihood Estimator assuming Poisson Model; 2) Likelihood Estimator assuming Geometric Model; 3) Likelihood Estimator assuming Poisson Model that drops out intercepts as well as subjects with only one measurement. Censoring process was assumed to be informative and dependent on both individual slopes and intercepts.

Simulation Parameters	Performance indicator $\times 10$	Estimator		
		1) Poisson	2) Geometric	3) Poisson that drops out intercepts and subjects with only one measurement
$\gamma_2 = -1.4 \quad \gamma_1 = -0.296$	Bias $_{\alpha}$	-0.074	-0.148	---
	Bias $_{\beta}$	0.073	0.164	0.185
	MSE(a) $_{\alpha}$	0.191	0.367	---
	MSE(a) $_{\beta}$	0.120	0.226	0.284
	MSE(b) $_{\alpha}$	0.195	0.406	---
	MSE(b) $_{\beta}$	0.034	0.098	0.124
$\gamma_2 = -5.2 \quad \gamma_1 = -0.296$	Bias $_{\alpha}$	-0.068	-0.145	---
	Bias $_{\beta}$	0.052	0.164	0.184
	MSE(a) $_{\alpha}$	0.180	0.373	---
	MSE(a) $_{\beta}$	0.139	0.180	0.210
	MSE(b) $_{\alpha}$	0.198	0.401	---
	MSE(b) $_{\beta}$	0.036	0.153	0.162
$\gamma_2 = -10.5 \quad \gamma_1 = -0.296$	Bias $_{\alpha}$	-0.066	-0.143	---
	Bias $_{\beta}$	0.054	0.141	0.183
	MSE(a) $_{\alpha}$	0.177	0.375	---
	MSE(a) $_{\beta}$	0.157	0.193	0.199
	MSE(b) $_{\alpha}$	0.198	0.400	---
	MSE(b) $_{\beta}$	0.038	0.148	0.163

* Datasets simulated from an underlying Poisson model with $\alpha=3.699$ $\beta=-0.138$, $\gamma_0 = 3.104$, $\sigma_{\alpha}^2 = 0.1445$, $\sigma_{\beta}^2 = 0.0187$, $\sigma_{\alpha\beta} = -0.0166$ and $\sigma_{\epsilon}^2 = 0.12$, with 2000 simulations.

Table 2

Summary * of Performance of Three Estimators: 1) Likelihood Estimator assuming Poisson Model; 2) Likelihood Estimator assuming Geometric Model; 3) Likelihood Estimator assuming Poisson Model that drops out intercepts as well as subjects with only one measurement. Censoring process was assumed to be informative and dependent on only individual intercepts (1), non-informative (2), and informative and dependent on only individual slopes (3).

Simulation Parameters	Performance indicator $\times 10$	Estimator		
		1) Poisson	2) Geometric	3) Poisson that drops out intercepts and subjects with only one measurement
1) $\gamma_2 = 0$ $\gamma_1 = -0.296$	Bias_α	-0.070	-0.128	---
	Bias_β	0.104	0.146	0.186
	MSE(a)_α	0.181	0.341	---
	MSE(a)_β	0.130	0.142	0.285
	MSE(b)_α	0.190	0.414	---
	MSE(b)_β	0.039	0.074	0.120
2) $\gamma_2 = 0$ $\gamma_1 = 0$	Bias_α	-0.079	-0.119	---
	Bias_β	0.028	0.086	0.182
	MSE(a)_α	0.187	0.331	---
	MSE(a)_β	0.079	0.116	0.283
	MSE(b)_α	0.186	0.379	---
	MSE(b)_β	0.021	0.030	0.113
3) $\gamma_2 = -5.2$ $\gamma_1 = 0$	Bias_α	-0.082	-0.123	---
	Bias_β	0.023	0.089	0.099
	MSE(a)_α	0.195	0.343	---
	MSE(a)_β	0.082	0.214	0.289
	MSE(b)_α	0.190	0.377	---
	MSE(b)_β	0.027	0.034	0.118

* Datasets simulated from an underlying Poisson model with $\alpha=3.699$ $\beta=-0.138$, $\gamma_0=3.104$, $\sigma_\alpha^2=0.1445$, $\sigma_\beta^2=0.0187$, $\sigma_{\alpha\beta}=-0.0166$ and $\sigma_\varepsilon^2=0.12$, with 2000 simulations.

Table 3

Summary * of Performance of Three Estimators: 1) Likelihood Estimator assuming Geometric Model; 2) Likelihood Estimator assuming Poisson Model; 3) Likelihood Estimator assuming Geometric Model that drops out intercepts as well as subjects with only one measurement. Censoring process was assumed to be informative and dependent on both individual slopes and intercepts.

Simulation Parameters	Performance indicator×10	Estimator		
		1) Geometric	2) Poisson	3) Geometric that drops out intercepts and subjects with only one measurement
$\gamma_2 = -1.4 \quad \gamma_1 = -3.115$	Bias _α	-0.084	-0.086	---
	Bias _β	0.040	0.045	0.053
	MSE(a) _α	0.103	0.103	---
	MSE(a) _β	0.126	0.341	0.360
	MSE(b) _α	0.099	0.128	---
	MSE(b) _β	0.051	0.197	0.206
$\gamma_2 = -5.2 \quad \gamma_1 = -3.115$	Bias _α	-0.083	-0.085	---
	Bias _β	0.061	0.097	0.099
	MSE(a) _α	0.101	0.102	---
	MSE(a) _β	0.167	0.283	0.299
	MSE(b) _α	0.098	0.120	---
	MSE(b) _β	0.125	0.189	0.193
$\gamma_2 = -10.5 \quad \gamma_1 = -3.115$	Bias _α	-0.084	-0.089	---
	Bias _β	0.092	0.133	0.155
	MSE(a) _α	0.104	0.113	---
	MSE(a) _β	0.173	0.180	0.187
	MSE(b) _α	0.096	0.125	---
	MSE(b) _β	0.054	0.099	0.118

* Datasets simulated from an underlying Geometric model with $\alpha=3.542$ $\beta=-0.131$ $\gamma_0=5.988$, $\sigma_\alpha^2=0.070$, $\sigma_\beta^2=0.033$, $\sigma_{\alpha\beta}=-0.029$ and $\sigma_\varepsilon^2=0.12$, with 2000 simulations.

Table 4

Summary * of Performance of Three Estimators: 1) Likelihood Estimator assuming Geometric Model; 2) Likelihood Estimator assuming Poisson Model; 3) Likelihood Estimator assuming Geometric Model that drops out intercepts as well as subjects with only one measurement. Censoring process was assumed to be informative and dependent on only individual intercepts (1), non-informative (2), and informative and dependent on only individual slopes (3).

Simulation Parameters	Performance indicator×10	Estimator		
		1) Geometric	2) Poisson	3) Geometric that drops out intercepts and subjects with only one measurement
1) $\gamma_2 = 0$ $\gamma_1 = -3.115$	Bias_α	-0.082	-0.083	---
	Bias_β	0.038	0.070	0.083
	MSE(a)_α	0.098	0.100	---
	MSE(a)_β	0.076	0.095	0.098
	MSE(b)_α	0.097	0.129	---
	MSE(b)_β	0.046	0.048	0.062
2) $\gamma_2 = 0$ $\gamma_1 = 0$	Bias_α	-0.078	-0.080	---
	Bias_β	0.033	0.064	0.068
	MSE(a)_α	0.084	0.096	---
	MSE(a)_β	0.071	0.092	0.095
	MSE(b)_α	0.093	0.118	---
	MSE(b)_β	0.041	0.047	0.084
3) $\gamma_2 = -5.2$ $\gamma_1 = 0$	Bias_α	-0.071	-0.083	---
	Bias_β	0.042	0.073	0.077
	MSE(a)_α	0.086	0.099	---
	MSE(a)_β	0.064	0.079	0.084
	MSE(b)_α	0.094	0.110	---
	MSE(b)_β	0.036	0.040	0.083

* Datasets simulated from an underlying Geometric model with $\alpha=3.542$ $\beta=-0.131$ $\gamma_0=5.988$, $\sigma_\alpha^2=0.070$, $\sigma_\beta^2=0.033$, $\sigma_{\alpha\beta}=-0.029$ and $\sigma_\varepsilon^2=0.12$, with 2000 simulations.