Published in final edited form as:

Biometrics. 2018 December ; 74(4): 1213–1222. doi:10.1111/biom.12911.

Semiparametric Regression Analysis of Interval-Censored Data with Informative Dropout

Fei Gao^{1,*}, Donglin Zeng^{2,**}, and Dan-Yu Lin^{2,***}

¹Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.

²Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

SUMMARY.

Interval-censored data arise when the event time of interest can only be ascertained through periodic examinations. In medical studies, subjects may not complete the examination schedule for reasons related to the event of interest. In this article, we develop a semiparametric approach to adjust for such informative dropout in regression analysis of interval-censored data. Specifically, we propose a broad class of joint models, under which the event time of interest follows a transformation model with a random effect and the dropout time follows a different transformation model but with the same random effect. We consider nonparametric maximum likelihood estimation and develop an EM algorithm that involves simple and stable calculations. We prove that the resulting estimators of the regression parameters are consistent, asymptotically normal, and asymptotically efficient with a covariance matrix that can be consistently estimated through profile likelihood. In addition, we show how to consistently estimate the survival function when dropout represents voluntary withdrawal and the cumulative incidence function when dropout is an unavoidable terminal event. Furthermore, we assess the performance of the proposed numerical and inferential procedures through extensive simulation studies. Finally, we provide an application to data on the incidence of diabetes from a major epidemiological cohort study.

Keywords

Joint models; Nonparametric likelihood; Random effects; Semiparametric efficiency; Terminal event; Transformation models

1. Introduction

Interval-censored data arise when the timing of an event is not known precisely but rather is known to lie within a time interval. Such data are frequently encountered in medical research, where the ascertainment of the disease of interest is made over a series of examination times. An example is the Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators, 1989), where subjects were examined for asymptomatic diseases, such as diabetes and hypertension, over five visits, with the first four each approximately

* feigao@uw.edu. ** dzeng@email.unc.edu. *** lin@bios.unc.edu.

three years apart and then a gap of about 15 years before the fifth visit, such that the disease was only known to occur within a broad time interval.

A number of methods have been developed for regression analysis of interval-censored data. In particular, nonparametric maximum likelihood estimation for the proportional odds, proportional hazards, and transformation models have been studied by Huang (1995), Huang (1996), and Zeng et al. (2016), respectively. Sieve estimation for the proportional odds and proportional hazards models has been suggested by Rossini and Tsiatis (1996), Huang and Rossini (1997), Shen (1998), and Cai and Betensky (2003). Rank-based estimation methods for linear transformation models have been proposed by Gu et al. (2005), Sun and Sun (2005), Zhang et al. (2005), and Zhang and Zhao (2013).

All aforementioned work assumes that the examination process is independent of the event of interest, possibly conditional on covariates. This assumption is often violated in chronic disease research because subjects may drop out of the study prematurely for health-related reasons. For example, in the ARIC study, a large number of subjects died before their last scheduled visit. When dropout is correlated with the event of interest, the existing methods may yield invalid inference. In the situation where dropout is caused by a terminal event, such as death, the existing methods, which fail to account for the fact that the event of interest cannot occur after the terminal event, will provide incorrect estimation of disease incidence even if dropout is independent of the event of interest.

In this article, we adjust for informative dropout through the use of a random effect. Specifically, we consider a broad class of joint models, under which the event time of interest follows a semiparametric transformation model with a random effect and the dropout time follows a different semiparametric transformation model but with the same random effect. The transformation models encompass the proportional hazards and proportional odds models. We study nonparametric maximum likelihood estimation for the joint models and develop a stable EM algorithm for its implementation. We establish the asymptotic properties of the resulting estimators, with different rates of convergence for the cumulative hazard functions of the event time of interest and the dropout time. In addition, we show how to predict the incidence for the event of interest when its occurrence is precluded by the development of a terminal event. Furthermore, we demonstrate the advantages of the proposed methods over the existing ones through realistic simulation studies. Finally, we provide a detailed illustration with data derived from the ARIC study.

2. Methods

2.1. Models and Likelihood

We consider a random sample of n subjects. For $i = 1, \dots, n$, let T_i denote the event time or failure time of interest, D_i the dropout time, and $X_i(\cdot)$ a p -vector of possibly time-dependent external covariates for the i th subject. We characterize the dependence between T_i and D_i through a random effect b_i , which is assumed to be normal with mean zero and variance σ^2 . Let X_i denote the entire history of the covariates. Conditional on b_i and X_i , the cumulative hazard functions for T_i and D_i follow the transformation models

$$\Lambda(t|b_i, X_i) = G \left\{ \int_0^t e^{\beta^T X_i(s) + b_i} d\Lambda(s) \right\} \quad (1)$$

and

$$A(t|b_i, X_i) = H \left\{ \int_0^t e^{\gamma^T X_i(s) + b_i} d\Lambda(s) \right\}, \quad (2)$$

respectively, where $G(\cdot)$ and $H(\cdot)$ are specific transformation functions, β and γ are unknown regression parameters, and $\Lambda(\cdot)$ and $A(\cdot)$ are arbitrary cumulative baseline hazard functions. For notational simplicity, we use the same X_i in models (1) and (2), although it is straightforward to use different sets of covariates.

REMARK 1.—Since there are only two possible events per subject, one shared random effect b_i is sufficient to capture the dependence and additional parameters would not be identifiable. The models with one shared random effect have been commonly adopted for two (right-censored) events, clustered sampling, and recurrent events (Oakes, 1989; Parner, 1998; Andersen et al., 2012, Chapter 9). For joint modeling of complex multivariate outcomes, such as longitudinal and survival outcomes or recurrent and terminal events, more than one random effect can be used (Zeng and Lin, 2007, 2009).

REMARK 2.—*To shed light on our modeling approach, we consider time-independent covariates, in which case models (1) and (2) can be expressed in the form of linear transformation models with a shared random effect*

$$\begin{cases} -\log\Lambda(T_i) = \beta^T X_i + b_i + \epsilon_{T,i} \\ -\log\Lambda(D_i) = \gamma^T X_i + b_i + \epsilon_{D,i} \end{cases}$$

where $\epsilon_{T,i}$ and $\epsilon_{D,i}$ are random errors with known distributions. This is a semiparametric version of the familiar linear mixed model for two correlated outcomes. Clearly, b_i characterizes the dependence between T_i and D_i in that the correlation between the two transformed event times is $\sigma^2 / \{(\sigma^2 + \sigma_T^2)(\sigma^2 + \sigma_D^2)\}^{1/2}$, where σ_T^2 and σ_D^2 are the variances of $\epsilon_{T,i}$ and $\epsilon_{D,i}$, respectively.

The transformation functions $G(\cdot)$ and $H(\cdot)$ include completely monotonic functions

$$G(x) = -\log \int_0^\infty e^{-xt} f_G(t) dt \quad (3)$$

and

$$H(x) = -\log \int_0^\infty e^{-xt} f_H(t) dt \quad (4)$$

where $f_G(\cdot)$ and $f_H(\cdot)$ are density functions with support on $[0, \infty)$. Particularly, the class of logarithmic transformations $r^{-1} \log(1 + rx)$ ($r > 0$) is generated by the gamma density function with mean 1 and variance r . The choice of $r=0$ or 1 yields the proportional hazards or proportional odds model, respectively.

REMARK 3.—We allow different transformation functions for the event time of interest and dropout time and let the data determine the best choices. It is natural to choose the proportional hazards models, which correspond to the identity transformation functions, especially when the sample size is not large enough to empirically determine the best transformation functions. The problem of interval-censored data with informative dropout has not been previously studied even under the proportional hazards models.

Suppose that the event of interest, such as diabetes, is asymptomatic, such that its occurrence can only be detected through periodic examinations. By contrast, dropout (e.g., death) can be observed exactly. There is a sequence of potential examination times for each subject. Obviously, no examination can occur after dropout. There exists noninformative censoring (e.g., end of the study), after which examination cannot occur either.

Specifically, let $0 < U_{i1} < \dots < U_{iM_i} < \infty$ denote the i th subject's potential examination times, which have finite support u with the least upper bound τ . Let C_i denote the noninformative censoring time on D_i such that we observe $Y_i \equiv \min(D_i, C_i)$ and $\delta_i \equiv I(D_i < C_i)$, where $I(\cdot)$ is the indicator function. The examination for the i th subject does not occur after Y_i . Examination typically does not occur at the time of dropout or the end of the study, such that Y_i is not equal to U_{ij} for any j . Thus, the failure time T_i is known to lie in the interval (L_i, R_i) , where $L_i = \max\{U_{im} : U_{im} < T_i, U_{im} < Y_i, m = 0, \dots, M_i\}$, $R_i = \min\{U_{im} : U_{im} \geq T_i, U_{im} < Y_i, m = 1, \dots, M_i\}$, and $U_{i0} = 0$. We let $R_i = \infty$ if the latter set is empty. If $Y_i < U_{i1}$, then no examination is performed and $(L_i, R_i) = (0, \infty)$. The observed data consist of O_i ($i = 1, \dots, n$) where $O_i = \{L_i, R_i, Y_i, \delta_i, X_i(\cdot)\}$.

Assume that $M_i, \{U_{im} : m = 1, \dots, M_i\}$, and C_i are independent of (T_i, D_i, b_i) conditional on X_i . The observed-data likelihood under models (1) and (2) is

$$\begin{aligned}
& \prod_{i=1}^n \int_{b_i} \left(\exp \left[-G \left\{ \int_0^{L_i} e^{\beta^T X_i(s) + b_i} d\Lambda(s) \right\} \right] \right. \\
& \quad \left. - \exp \left[-G \left\{ \int_0^{R_i} e^{\beta^T X_i(s) + b_i} d\Lambda(s) \right\} \right] \right) \\
& \quad \times \left[e^{\gamma^T X_i(Y_i) + b_i} \int_0^{Y_i} e^{\gamma^T X_i(s) + b_i} dA(s) \right]^{\Delta_i} \\
& \quad \times \exp \left[-H \left\{ \int_0^{Y_i} e^{\gamma^T X_i(s) + b_i} dA(s) \right\} \right] \phi(b_i; \sigma^2) db_i
\end{aligned}$$

where $g'(\cdot)$ denotes the derivative of the function $g(\cdot)$, $\phi(b; \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp(-b^2/2\sigma^2)$, and we define $\exp \left[-G \left\{ \int_0^\infty e^{\beta^T X_i(s) + b_i} d\Lambda(s) \right\} \right] = 0$.

2.2. Estimation Procedures

We adopt the nonparametric maximum likelihood approach, under which the estimators for the cumulative baseline hazard functions Λ and A are step functions with jumps at the unique endpoints of the intervals, $0 < t_1 < \dots < t_{m1} < \infty$, and at the uncensored dropout times, $0 < s_1 < \dots < s_{m2} < \infty$, respectively, where m_1 and m_2 are the total numbers of potential jump points. We denote the step sizes for Λ as $\lambda_1, \dots, \lambda_{m1}$ and the step sizes for A as $\alpha_1, \dots, \alpha_{m2}$. Write $\theta = (\beta, \gamma, \sigma^2)$ and $\mathcal{A} = (\Lambda, A)$. We maximize the objective function

$$L_n(\theta, \mathcal{A}) \equiv \prod_{i=1}^n \int_{b_i} L_i^{(1)}(b_i; \beta, \Lambda) L_i^{(2)}(b_i; \gamma, A) \phi(b_i; \sigma^2) db_i$$

over $\beta, \gamma, \sigma^2, (\lambda_1, \dots, \lambda_{m1})$, and $(\alpha_1, \dots, \alpha_{m2})$, where

$$\begin{aligned}
L_i^{(1)}(b_i; \beta, \Lambda) &= \exp \left[-G \left(\sum_{t_l \leq L_i} e^{\beta^T X_{il} + b_i} \lambda_l \right) \right] \\
&\quad - \exp \left[-G \left(\sum_{t_l \leq R_i} e^{\beta^T X_{il} + b_i} \lambda_l \right) \right]
\end{aligned}$$

and

$$L_i^{(2)}(b_i; \gamma, A) = \left[A\{Y_i\} e^{\gamma^T X_i(Y_i) + b_i} H' \left(\sum_{s_l \leq Y_i} e^{\gamma^T X_{il}^* + b_i} \alpha_l \right) \right]^{\Delta_i} \exp \left\{ -H \left(\sum_{s_l \leq Y_i} e^{\gamma^T X_{il}^* + b_i} \alpha_l \right) \right\}$$

with $X_{il} = X_i(t_l)$ for $l = 1, \dots, m_1$, $X_{il}^* = X_i(s_l)$ for $l = 1, \dots, m_2$, and $A\{Y_i\}$ being the jump size of A at Y_i .

Direct maximization of $L_n(\theta, A)$ is difficult due to the lack of analytical expressions for the parameters $\lambda_1, \dots, \lambda_{m_1}$ and $\alpha_1, \dots, \alpha_{m_2}$. We introduce some latent random variables to form a likelihood function equivalent to $L_n(\theta, A)$ such that the maximization can be carried out by a simple EM algorithm. First, we introduce two latent random variables ξ_i and ψ_i with density functions $f_G(\cdot)$ and $f_H(\cdot)$ given in equations (3) and (4), respectively. We then introduce independent Poisson random variables $W_{il} (l = 1, \dots, m_1, t_l \leq R_i^*)$ with means $\lambda_i \xi_i \exp(\beta^T X_{il} + b_i)$, where $R_i^* = L_i I(R_i = \infty) + R_i I(R_i < \infty)$. Conditional on (ξ_i, b_i) , the likelihood function of $\{W_{il}; l = 1, \dots, m_1, t_l \leq R_i^*\}$ is

$$\prod_{l=1, t_l \leq R_i^*}^{m_1} \left\{ \frac{1}{W_{il}!} \left(\lambda_i \xi_i e^{\beta^T X_{il} + b_i} \right)^{W_{il}} \exp \left(-\lambda_i \xi_i e^{\beta^T X_{il} + b_i} \right) \right\}.$$

Let $N_{1i} = \sum_{t_l \leq L_i} W_{il}$ and $N_{2i} = I(R_i < \infty) \sum_{L_i < t_l \leq R_i} W_{il}$. Suppose that we observe $N_{1i} = 0$ and $N_{2i} < 0$. The observed-data likelihood for $N_{1i} = 0$ and $N_{2i} > 0$ given ξ_i and b_i is equal to

$$g_{i1}(\xi_i, b_i) \equiv \exp \left(-\xi_i \sum_{t_l \leq L_i} e^{\beta^T X_{il} + b_i} \lambda_l \right) - I(R_i < \infty) \exp \left(-\xi_i \sum_{t_l \leq R_i} e^{\beta^T X_{il} + b_i} \lambda_l \right).$$

In addition, the observed-data likelihood for (Y_i, Δ_i) given ψ_i and b_i is

$$g_{i2}(\psi_i, b_i) \equiv \left[\psi_i A'(Y_i) e^{\gamma^T X_i(Y_i) + b_i} \right]^{\Delta_i} \exp \left\{ -\psi_i \int_0^{Y_i} e^{\gamma^T X_i(s) + b_i} dA(s) \right\}.$$

Therefore, $L_i^{(1)}(b_i; \beta, \Lambda) = \int_{\xi_i} g_{i1}(\xi_i, b_i) f_G(\xi_i) d\xi_i$, and $L_i^{(2)}(b_i; \gamma, A) = \int_{\psi_i} g_{i2}(\psi_i, b_i) f_H(\psi_i) d\psi_i$. In other words, $L_n(\theta, \mathcal{A})$ can be viewed as the observed-data likelihood for $\{N_{1i} = 0, N_{2i} > 0, Y_i, \Delta_i\}$ with $(W_{il}, \xi_i, \psi_i, b_i)$ ($l = 1, \dots, m_1, t_l \leq R_i^*$) as latent variables. Based on the foregoing results, we propose an EM algorithm treating $(W_{il}, \xi_i, \psi_i, b_i)$ ($i = 1, \dots, n; l = 1, \dots, m_1, t_l \leq R_i^*$) as complete data.

In the M-step, we maximize the conditional expectation of the complete-data log-likelihood given the observed data so as to update the parameters. Specifically, we update β by solving the equation

$$\sum_{i=1}^n \sum_{l=1}^{m_1} \hat{E}(W_{il}) I(t_l \leq R_i^*) \left[X_{il} - \frac{\sum_{j=1}^n X_{jl} I(t_l \leq R_j^*) \hat{E}\{\xi_j \exp(\beta^T X_{jl} + b_j)\}}{\sum_{j=1}^n I(t_l \leq R_j^*) \hat{E}\{\xi_j \exp(\beta^T X_{jl} + b_j)\}} \right] = 0,$$

and we update Λ by

$$\lambda_l = \frac{\sum_{i=1}^n I(t_l \leq R_i^*) \hat{E}(W_{il})}{\sum_{i=1}^n I(t_l \leq R_i^*) \hat{E}\{\xi_i \exp(\beta^T X_{il} + b_i)\}}, l = 1, \dots, m_1,$$

where $\hat{E}(\cdot)$ denotes the conditional expectation given the observed data

$\tilde{\mathcal{O}}_i \equiv \{N_{1i} = 0, N_{2i} > 0, Y_i, \Delta_i, X_i(\cdot)\}$ ($i = 1, \dots, n$). In addition, we update γ by solving the equation

$$\sum_{i=1}^n \Delta_i \left(X_i(Y_i) - \frac{\sum_{j=1}^n I(Y_j \geq Y_i) X_j(Y_i) \hat{E}\{\psi_j \exp(\gamma^T X_j(Y_i) + b_j)\}}{\sum_{j=1}^n I(Y_j \geq Y_i) \hat{E}\{\psi_j \exp(\gamma^T X_j(Y_i) + b_j)\}} \right) = 0,$$

and we update A by

$$\alpha_l = \frac{\sum_{i=1}^n \Delta_i I(Y_i = s_l)}{\sum_{i=1}^n I(Y_i \geq s_l) \hat{E}\{\psi_i \exp(\gamma^T X_{il}^* + b_i)\}}, l = 1, \dots, m_2.$$

Finally, we update σ^2 by $\sigma^2 = n^{-1} \sum_{i=1}^n \hat{E}(b_i^2)$.

In the E-step, we evaluate the conditional expectation of W_{il} ($l = 1, \dots, m_1, t_l \leq R_i^*$) and the other terms of ξ_i , ψ_i , and b_i given the observed data $\tilde{\mathcal{O}}_i$ for $i = 1, \dots, n$. Specifically, the conditional expectation of W_{il} ($l = 1, \dots, m, t_l \leq R_i^*$) given $\tilde{\mathcal{O}}_i$, ξ_i , and b_i is

$$I(L_i < t_l \leq R_i < \infty) \frac{\lambda_l^{\xi_i} \exp(\beta^T X_{il} + b_i)}{1 - \exp\left(-\sum_{L_i < t_{l'} < R_i} \lambda_{l'}^{\xi_i} e^{\beta^T X_{il'} + b_i}\right)}.$$

Note that the joint density of (ξ_i, ψ_i, b_i) given $\tilde{\theta}_i$ is proportional to

$g_{i1}(\xi_i, b_i) f_G(\xi_i) g_{i2}(\psi_i, b_i) \times f_H(\psi_i) \phi(b_i; \sigma^2)$. We evaluate the conditional expectation of W_{il} and the other terms through numerical integration over ξ_i , ψ_i and b_i with Gauss–Laguerre and Gauss–Hermite quadratures.

We iterate between the E-step and the M-step until the sum of the absolute differences of the parameter values at two successive iterations is less than a certain threshold, say, 10^{-3} . We denote the final estimators for θ and $\mathcal{A}$ as $\hat{\theta} \equiv (\hat{\beta}, \hat{\gamma}, \hat{\sigma}^2)$ and $\hat{\mathcal{A}} \equiv (\hat{\Lambda}, \hat{\mathcal{A}})$. The survival function for the failure time of interest, $P(T \geq t|X)$, can be estimated by

$$\int_b \exp\left[-G\left\{\sum_{t_l \leq t} e^{\hat{\beta}^T X(t_l) + b_{\hat{\lambda}_l}}\right\}\right] \phi(b; \hat{\sigma}^2) db.$$

REMARK 4.—The proposed EM algorithm has several desirable features. First, large-scale optimization is avoided as jump sizes are updated explicitly in the M-step. Second, the regression parameters are updated by solving estimating equations similar to the Poisson regression score equations and the partial likelihood score equations via one-step Newton–Raphson. Finally, the E-step involves only two-dimensional numerical integration.

If dropout is a terminal event, which cannot be avoided, then we have a semi-competing risks set-up (Fine et al., 2001) in that the occurrence of the terminal event precludes the development of the event of interest but not vice versa. It is more meaningful to consider the cumulative incidence function for the failure time of interest

$$\begin{aligned} & P(T \leq t, T \leq D|X) \\ &= P(T \leq t \leq D|X) + P(T \leq D < t|X) \\ &= \int_b \left(1 - \exp\left[-G\left\{\int_0^t e^{\beta^T X(u) + b_{d\Lambda(u)}}\right\}\right]\right) \\ &\quad \exp\left[-H\left\{\int_0^t e^{\gamma^T X(u) + b_{dA(u)}}\right\}\right] \phi(b; \sigma^2) db \\ &\quad + \int_b \left[\int_0^t \left\{e^{\gamma^T X(s) + b_{H'}} \left\{\int_0^s e^{\gamma^T X(u) + b_{dA(u)}}\right\}\right.\right. \\ &\quad \left.\left.\exp\left[-H\left\{\int_0^s e^{\gamma^T X(u) + b_{dA(u)}}\right\}\right]\right.\right. \\ &\quad \left.\left.\times \left(1 - \exp\left[-G\left\{\int_0^s e^{\beta^T X(u) + b_{d\Lambda(u)}}\right\}\right]\right)\right] d\Lambda(s)\right] \phi(b; \sigma^2) db. \end{aligned}$$

We estimate this quantity by

$$\begin{aligned} & \int_b \left(1 - \exp \left[-G \left(\sum_{t_l \leq t} e^{\hat{\beta}^T X(t_l) + b \hat{\lambda}_l} \right) \right] \right) \\ & \exp \left[-H \left(\sum_{s_l \leq t} e^{\hat{\gamma}^T X(s_l) + b \hat{\alpha}_l} \right) \right] \phi(b; \sigma^2) db \\ & + \int_b \left(\sum_{s_l \leq t} \left\{ \hat{\alpha}_l e^{\hat{\gamma}^T X(s_l) + b} \right\}_{H'} \left(\sum_{t_{l'} \leq t} e^{\hat{\gamma}^T X(t_{l'}) + b \hat{\alpha}_{l'}} \right) \right. \\ & \left. \exp \left[-H \left(\sum_{t_{l'} \leq t} e^{\hat{\gamma}^T X(t_{l'}) + b \hat{\alpha}_{l'}} \right) \right] \right) \\ & \times \left(1 - \exp \left[-G \left(\sum_{t_{l'} \leq s_l} e^{\hat{\beta}^T X(t_{l'}) + b \hat{\lambda}_{l'}} \right) \right] \right) \phi(b; \sigma^2) db, \end{aligned}$$

where the integral is evaluated by numerical integration with Gauss–Hermite quadratures, and $\hat{\lambda}_l$ and $\hat{\alpha}_l$ are the estimators of λ_l and α_l , respectively.

We have implicitly assumed that the transformation functions are known. In practice, we consider a variety of transformation models and choose the one that best fits the data according to, say, the Akaike information criterion.

2.3. Inference Procedures

In Web Appendix A, we show that the estimators $\hat{\theta}$ and $\hat{\mathcal{A}}$ are consistent, with different rates of convergence ($n^{1/3}$ and $n^{1/2}$, respectively) for $\hat{\lambda}$ and $\hat{\alpha}$. In addition, the estimator $\hat{\theta}$ is asymptotically normal with a limiting covariance matrix that attains the semiparametric efficiency bound.

To estimate the covariance matrix of $\hat{\theta}$, we define the profile likelihood function

$$pl_n(\theta) = \max_{\mathcal{A} \in C_1 \times C_2} \log L_n(\theta, \mathcal{A}),$$

where C_1 is the set of step functions with non-negative jumps at t_l ($l = 1, \dots, m_1$), and C_2 is the set of step functions with non-negative jumps at s_l ($l = 1, \dots, m_2$). Let $pli(\theta)$ denote the i th subject's contribution to $pli(\theta)$, that is, $pl_i(\theta) = l_i(\theta, \hat{\mathcal{A}}_\theta)$, where $l_i(\theta, \mathcal{A})$ is the log-likelihood function for the i th subject, and $\hat{\mathcal{A}}_\theta = \arg \max \log L_n(\theta, \mathcal{A})$. We then estimate the covariance matrix of $\hat{\theta}$ by the inverse of the matrix

$$\sum_{i=1}^n \begin{pmatrix} \frac{pl_i(\hat{\theta} + h_n e_1) - pl_i(\hat{\theta})}{h_n} \\ \vdots \\ \frac{pl_i(\hat{\theta} + h_n e_{2p+1}) - pl_i(\hat{\theta})}{h_n} \end{pmatrix}^{\otimes 2},$$

where e_j is the j th canonical vector in $\mathbb{R}^{2p+1}$, $a^{\otimes 2} = aa^T$, and h_n is a constant of order $n^{-1/2}$. To evaluate the profile likelihood, we use the EM algorithm of Section 2.2 but fix the value of θ and only update Λ and A in the M-step.

REMARK 5.—*In the conventional profile-likelihood approach (Murphy and van der Vaart, 2000), the covariance matrix of $\hat{\theta}$ is estimated by the negative inverse of the Hessian matrix of $pl_n(\theta)$ at $\hat{\theta}$ which is determined by the second-order numerical differences. The Hessian matrix may be negative definite, especially in small samples. By contrast, we use the empirical covariance matrix of the gradient of $pl_n(\theta)$ based on the first-order numerical differences, which approximates the efficient score function for θ . The calculation is quicker, and the resulting covariance matrix estimator is guaranteed to be positive semidefinite.*

3. Simulation Studies

We conducted two sets of simulation studies to assess the performance of the proposed methods. In the first set, we considered one time-independent covariate 1) and one time-dependent covariate $X_1 \sim \text{Unif}(0, 1)$, and one time-independent covariate $X_2(t) = \tilde{B}_1 I(t \leq V) + \tilde{B}_2 I(t \leq V) + \tilde{B}_2 I(t > V)$, where $\tilde{B}_1$ and $\tilde{B}_2$ are independent Bernoulli(0.5), $V \sim \text{Unif}(0, \tau)$, and $\tau = 4$. We considered logarithmic transformation functions $G(x) = r_G^{-1} \log(1 + r_G x)$ and $H(x) = r_H^{-1} \log(1 + r_H x)$. We set

$$\beta \equiv (\beta_1, \beta_2)^T = (0.5, 0.4)^T, \gamma \equiv (\gamma_1, \gamma_2)^T = (0.5, 0.2)^T, \sigma^2 = 0.5, \Lambda(t) = 0.5t, \text{ and } A(t) = \log(1 + 0.5t).$$

We generated the potential examination times $U_m \sim U_m - 1 + 0.1 + \text{Unif}(0, \tau/5)$ with $U_0 = 0$ and the censoring time C from $\text{Unif}(2\tau/3, \tau)$. The number of actual examinations is approximately 2.4 per subject. The event time of interest is left-censored for 19% of the subjects, interval-censored for 28% of the subjects, and right-censored for 53% of the subjects. We set $n = 100, 200$, or 400 and used 10,000 replicates. For each dataset, we applied the proposed EM algorithm by setting the initial values of β and γ to $\mathbf{0}$, the initial value of σ^2 to 1 and the initial values of λ_j and α_j to $1/m_j$ and $1/m_2$, respectively. We used 20 quadrature points for integration with respect to each random effect and set the convergence threshold to 10^{-3} . The variance estimators were obtained with $h_n = n^{-1/2}$.

Table 1 summarizes the results on the estimation of β , γ , and σ^2 for different values of n , r_G , and r_H . The biases for all parameter estimators are small and decrease as n increases. For large n , the variance estimator for $\hat{\beta}$ and $\hat{\gamma}$ are accurate especially for large n . The variance

estimators for $\hat{\sigma}^2$ is accurate, under $r_G = r_H = 0$ but slightly overestimates the true variabilities under $r_G = r_H = 1$. The confidence intervals for β , γ , and σ^2 have reasonable coverage probabilities.

We also evaluated the method of Zeng et al. (2016), which does not account for informative dropout. The results for this naive method are shown in Table 2. The estimator for β is biased, and the coverage probability of the corresponding confidence interval is poor.

Figure 1a shows the estimation of the baseline survival function for the event time of interest when dropout is regarded as voluntary patient withdrawal. The proposed estimator is virtually unbiased, whereas the naive method (Zeng et al., 2016) overestimates the survival function. Figure 1b shows the estimation of the baseline cumulative incidence function for the event time of interest when dropout is treated as a terminal event. The proposed estimator is again virtually unbiased; the naive estimator has severe positive bias since it does not acknowledge the fact that the event of interest cannot occur after the terminal event.

In the second set of simulation studies, we set $\beta = (-0.5, 0)^T$, $\gamma = (0, 0.4)^T$, and $\sigma^2 = 1$ and considered $r_G = r_H = 0, 0.5$, or 1 . The results are presented in Web Appendix B. The basic conclusions are the same as those of the first set.

4. ARIC Study

ARIC is a prospective epidemiological study conducted in four U.S. communities: Forsyth County, NC; Jackson, MS; suburbs of Minneapolis, MN; and Washington County, MD (The ARIC investigators, 1989). One important objective is to investigate risk factors for diabetes. A total of 14,751 Caucasian and African-American participants underwent a baseline examination between 1987 and 1989 and were scheduled for four subsequent examinations to take place in 1990–1992, 1993–1995, 1996–1998, and 2011–2013. Diabetes status (defined as fasting glucose ≥ 126 mg/dL, non-fasting glucose ≥ 200 mg/dL, self-reported physician diagnosis of diabetes, or use of diabetic medication) was determined at each examination.

We related the incidence of diabetes and death to race, gender, community, and five baseline risk factors: age, body mass index, glucose level, systolic blood pressure, and diastolic blood pressure. We excluded 1933 subjects with prevalent diabetes or unknown diabetes status at baseline and 13 subjects with missing baseline covariate values to obtain a total of 12,805 subjects. Among those subjects, 11,686 (91.3%), 10,557 (82.4%), 9533 (74.4%), and 5035 (39.3%) completed the second, third, fourth, and fifth visits, respectively. As shown in Web Figure 3, there are sufficient overlaps of the visit times for us to study diabetes onset from year 2 to 12 and from year 22 to 27. A total of 2492 (19.5%) subjects developed diabetes during the study, and 4363 (34.1%) subjects died before the end of the study.

We fit models (1) and (2) with logarithmic transformation functions indexed by parameters r_G and r_H for diabetes and death, respectively. The likelihood is maximized at $r_G = 2.3$ and $r_H = 0$, which is the combination that would be selected by the Akaike information criterion. For easy interpretation, we set $r_G = 1$ and $r_H = 0$.

Table 3 shows the estimation results for the proportional hazards models for both events ($r_G = r_H = 0$), the proportional odds models for both events ($r_G = r_H = 1$), and the combination of the proportional odds model for diabetes and the proportional hazards model for death ($r_G = 1, r_H = 0$). The log-likelihood values are approximately -48724.5 , -48707.3 , and -48681.1 for the three combinations of transformation parameters. The variance component σ^2 was estimated to be 0.561, 0.681, and 0.530 with standard errors 0.063, 0.096, and 0.070, respectively, for the three combinations of transformation parameters, indicating strong dependence between diabetes and death. Under all considered models, an African–American individual has a higher risk of diabetes than a Caucasian individual. In addition, higher baseline body mass index, glucose level, and systolic blood pressure are associated with increased risk for diabetes. These findings are consistent with the current literature on diabetes (Harris et al., 1987; DeFronzo et al., 2004).

The results from the naive method, which are shown in Table 4, are considerably different from ours. In particular, the naive method identifies a negative association between age and risk of diabetes, which contradicts the established positive association in the literature (Harris et al., 1987). The proposed method adjusting for death finds no significant negative association. The relationship between age and risk of diabetes identified by the naive method is likely a spurious finding that reflects the strong correlation between age and death.

Figure 2 compares the estimated cumulative incidence functions for an African–American male versus a Caucasian male with the same values of other risk factors. The risk of diabetes is considerably higher for the African–American individual than the Caucasian individual under all considered models, with appreciably different estimates between the proportional hazards and proportional odds models. The estimated probabilities from the proposed method are lower in the tail than their naive counterparts, especially under the proportional odds model, highlighting the importance of adjusting for death.

5. Discussion

In this article, we study efficient nonparametric maximum likelihood estimation of joint models for interval-censored data with informative dropout. We establish the asymptotic properties of the estimators through innovative use of modern empirical process theory. In the proofs, separate treatments are given to the estimators of the cumulative baseline hazard functions for the event of interest and dropout. We avoid the assumption of Zeng et al. (2016) that a subset of study subjects are examined at the end of the study by carefully evaluating the bracket covering number for a class of functions that involves unbounded Λ .

We applied our methods to data derived from the ARIC study, where diabetes is the event of interest and death is the terminal event. In the ARIC study, there are other outcomes of interest that are either interval-censored (e.g., hypertension, peripheral artery disease) or right-censored (e.g., myocardial infarction, stroke). The proposed framework can be extended to incorporate multiple interval-censored events and multiple right-censored events and thereby analyze an enriched version of data from the ARIC study.

The class of transformation models is very broad and thus allows accurate prediction in a variety of situations. In practice, one would need to determine which model best fits the data. One strategy is to use the Akaike information criterion to select the best transformations, as we did for the ARIC study. It would be worthwhile to develop additional methods for model selection and model checking.

We have assumed that the transformation functions are correctly specified. If the transformation functions or other aspects of the models are incorrectly specified, then the proposed parameter estimators will converge to constants that minimize the Kullback–Leibler distance. Robust variance estimators can be constructed on the basis of the efficient score functions. A thorough investigation of model misspecification is beyond the scope of this article.

In some applications, the examination times are directly related to the event of interest, instead of through dropout. This may be the case if patients tend to visit their doctors more frequently when they are not feeling well. Zhang et al. (2005), Chen et al. (2012, 2014), and Ma et al. (2015) studied this problem for current status data (Huang, 1996) by assuming a frailty model or copula structure for the event time of interest and the examination time. Zhang et al. (2007) considered the case of two examination times and modeled the first examination time, the gap time, and the event time of interest through a proportional hazards frailty model. Zhao et al. (2015) considered the same type of data and assumed a copula model for the event time of interest and the gap time. Wang et al. (2016) considered an arbitrary number of examination times and assumed a shared frailty model. All of these methods require parametric assumptions or approximations for the cumulative baseline hazard functions.

We can extend our approach to the aforementioned settings. For current status data, our algorithm can be directly applied by replacing the dropout time with the examination time in the likelihood. For the case with two examination times, we can jointly model the failure time, the first examination time, and the gap time using transformation models with random effects. For an arbitrary number of examination times, we can model the intensity of the examination process using a transformation model with frailty. The proposed EM algorithm can be modified accordingly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health awards R01GM047845, R01AI029168, R01CA082659, and P01CA142538. The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions.

References

- Andersen PK, Borgan O, Gill RD, and Keiding N (2012). *Statistical Models Based on Counting Processes* New York: Springer.
- Cai T and Betensky RA (2003). Hazard regression for intervalcensored data with penalized spline. *Biometrics* 59, 570–579. [PubMed: 14601758]
- Chen CM, Lu TFC, Chen MH, and Hsu CM (2012). Semiparametric transformation models for current status data with informative censoring. *Biometrical Journal* 54, 641–656. [PubMed: 22886604]
- Chen CM, Wei JCC, Hsu CM, and Lee MY (2014). Regression analysis of multivariate current status data with dependent censoring: application to ankylosing spondylitis data. *Statistics in Medicine* 33, 772–785. [PubMed: 24122926]
- DeFronzo RA, Ferrannini E, Zimmet P, and Alberti G (2004). *International Textbook of Diabetes Mellitus* Chichester, U.K: John Wiley and Sons.
- Fine JP, Jiang H, and Chappell R (2001). On semi-competing risks data. *Biometrika* 88, 907–919.
- Gu MG, Sun L, and Zuo G (2005). A baseline-free procedure for transformation models under interval censorship. *Lifetime Data Analysis* 11, 473–488. [PubMed: 16328572]
- Harris MI, Hadden WC, Knowler WC, and Bennett PH (1987). Prevalence of diabetes and impaired glucose tolerance and plasma glucose levels in US population aged 20–74 yr. *Diabetes* 36, 523–534. [PubMed: 3817306]
- Huang J (1995). Maximum likelihood estimation for proportional odds regression model with current status data. *Analysis of Censored Data (IMS Lecture Notes Monograph Series)* 27, 129–145.
- Huang J (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* 24, 540–568.
- Huang J and Rossini A (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association* 92, 960–967.
- Ma L, Hu T, and Sun J (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* 102, 731–738.
- Murphy SA and van der Vaart AW (2000). On profile likelihood. *Journal of the American Statistical Association* 95, 449–465.
- Oakes D (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84, 487–493.
- Parner E (1998). Asymptotic theory for the correlated gammafrailty model. *The Annals of Statistics* 26, 183–214.
- Rossini A and Tsiatis A (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* 91, 713–721.
- Shen X (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* 85, 165–177.
- Sun J and Sun L (2005). Semiparametric linear transformation models for current status data. *Canadian Journal of Statistics* 33, 85–96.
- The ARIC Investigators (1989). The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. *American Journal of Epidemiology* 129, 687–702. [PubMed: 2646917]
- Wang P, Zhao H, and Sun J (2016). Regression analysis of case K interval-censored failure time data in the presence of informative censoring. *Biometrics* 72, 1103–1112. [PubMed: 27123560]
- Zeng D and Lin D (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *Journal of the Royal Statistical Society, Series B* 69, 507–564.
- Zeng D and Lin D (2009). Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics* 65, 746–752. [PubMed: 18945267]
- Zeng D, Mao L, and Lin D (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* 103, 253–271. [PubMed: 27279656]
- Zhang Z, Sun J, and Sun L (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine* 24, 1399–1407. [PubMed: 15565735]
- Zhang Z, Sun L, Sun J, and Finkelstein DM (2007). Regression analysis of failure time data with informative interval censoring. *Statistics in Medicine* 26, 2533–2546. [PubMed: 17072823]

- Zhang Z, Sun L, Zhao X, and Sun J (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics* 33, 61–70.
- Zhang Z and Zhao Y (2013). Empirical likelihood for linear transformation models with interval-censored failure time data. *Journal of Multivariate Analysis* 116, 398–409.
- Zhao S, Hu T, Ma L, Wang P, and Sun J (2015). Regression analysis of informative current status data with the additive hazards model. *Lifetime Data Analysis* 21, 241–258. [PubMed: 25078350]

(a) Survival function

(b) Cumulative incidence function

Figure 1.

Estimation of (a) the baseline survival function and (b) the baseline cumulative incidence function. The solid, dashed, and dotted curves pertain, respectively, to the true value, mean estimate from the proposed method, and mean estimate from the naive method. This figure appears in color in the electronic version of this article.

Figure 2.

Estimation of cumulative incidence functions for (a) an African-American male versus (b) a Caucasian male residing in Forsyth County, NC, aged 54 years, body mass index 27kg/m^2 , glucose value 98mg/dl , systolic blood pressure 118mmHg , and diastolic blood pressure 73mmHg . The solid and dashed curves pertain to the proportional hazards and proportional odds models, respectively, from the proposed method, where the dropout time is modeled by the proportional hazards model. The dotted and dash-dotted curves pertain to the proportional hazards and proportional odds models, respectively, from the naive method. This figure appears in color in the electronic version of this article.

Table 1

Summary statistics for the simulation studies on the proposed estimators

	$r_G = r_H = 0$						$r_G = r_H = 1$					
	Bias	SE	SEE	CP	Bias	SE	Bias	SE	SEE	CP	Bias	CP
n=100												
β_1	0.027	0.728	0.679	0.95	0.030	0.921	0.872	0.94				
β_2	0.028	0.422	0.384	0.94	0.037	0.533	0.482	0.94				
γ_1	0.002	0.544	0.530	0.96	-0.012	0.723	0.728	0.97				
γ_2	0.009	0.288	0.290	0.96	0.000	0.381	0.384	0.96				
σ^2	-0.019	0.555	0.523	0.97	-0.112	0.781	0.909	0.96				
n=200												
β_1	0.006	0.482	0.460	0.94	0.009	.623	0.596	0.94				
β_2	0.016	0.277	0.262	0.94	0.021	0.356	0.329	0.93				
γ_1	0.002	0.368	0.365	0.96	0.010	0.498	0.501	0.96				
γ_2	0.000	0.198	0.200	0.96	0.004	0.260	0.262	0.96				
σ^2	-0.002	0.367	0.357	0.96	-0.044	0.556	0.635	0.95				
n=400												
β_1	0.000	0.329	0.318	0.94	0.011	0.429	0.412	0.94				
β_2	0.007	0.187	0.182	0.95	0.012	0.244	0.228	0.94				
γ_1	0.005	0.255	0.254	0.95	0.002	0.354	0.349	0.95				
γ_2	0.003	0.140	0.139	0.95	0.003	0.182	0.182	0.95				
σ^2	0.002	0.250	0.245	0.96	-0.015	0.401	0.443	0.95				

Note: Bias, SE, SEE, and CP stand, respectively, for the median bias, empirical standard error, median standard error estimator, and empirical coverage percentage of the 95% confidence interval. For σ^2 , the confidence interval is based on the log transformation.

Table 2

Summary statistics for the simulation studies on the naive method

	$r_G = r_H = 0$				$r_G = r_H = 1$			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP
n=100								
β_1	-0.125	0.600	0.560	0.93	-0.054	0.829	0.773	0.93
β_2	-0.058	0.355	0.324	0.93	-0.015	0.483	0.433	0.92
n=200								
β_1	-0.143	0.400	0.383	0.93	-0.074	0.559	0.536	0.93
β_2	-0.067	0.236	0.224	0.93	-0.025	0.324	0.299	0.93
n=400								
β_1	-0.148	0.274	0.267	0.91	-0.069	0.387	0.375	0.93
β_2	-0.075	0.160	0.156	0.92	-0.033	0.222	0.209	0.93

Note: Bias, SE, SEE, and CP stand, respectively, for the median bias, empirical standard error, median standard error estimator, and empirical coverage percentage of the 95% confidence interval.

Table 3

Regression analysis for diabetes in the ARIC study with adjustments for death

r_G	r_H	Covariate	Diabetes			Death		
			Est	Std. Err	p-value	Est	Std. Err	p-value
0	0	Jackson	-0.185	0.097	0.057	0.039	0.096	0.684
		Minneapolis Suburbs	-0.424	0.069	< 10 ⁻⁴	-0.032	0.052	0.534
		Washington County	0.101	0.065	0.123	0.088	0.050	0.080
		Age	-0.004	0.004	0.358	0.104	0.004	< 10 ⁻⁴
		Male	-0.010	0.046	0.831	0.565	0.036	< 10 ⁻⁴
		White	-0.485	0.110	< 10 ⁻⁴	-0.503	0.100	< 10 ⁻⁴
		Body mass index	0.075	0.004	< 10 ⁻⁴	-0.017	0.018	0.363
		Glucose	0.102	0.003	< 10 ⁻⁴	0.055	0.018	0.002
		Systolic blood pressure	0.006	0.002	0.001	0.313	0.024	< 10 ⁻⁴
		Diastolic blood pressure	-0.001	0.003	0.793	-0.162	0.025	< 10 ⁻⁴
1	1	Jackson	-0.242	0.088	0.006	0.096	0.087	0.269
		Minneapolis Suburbs	-0.510	0.082	< 10 ⁻⁴	-0.034	0.063	0.587
		Washington County	0.118	0.078	0.132	0.121	0.061	0.050
		Age	-0.008	0.005	0.156	0.124	0.004	< 10 ⁻⁴
		Male	-0.033	0.055	0.547	0.680	0.045	< 10 ⁻⁴
		White	-0.648	0.110	< 10 ⁻⁴	-0.604	0.101	< 10 ⁻⁴
		Body mass index	0.096	0.006	< 10 ⁻⁴	-0.005	0.004	0.265
		Glucose	0.124	0.003	< 10 ⁻⁴	0.007	0.002	0.003
		Systolic blood pressure	0.007	0.002	0.001	0.020	0.002	< 10 ⁻⁴
		Diastolic blood pressure	-0.001	0.004	0.891	-0.017	0.003	< 10 ⁻⁴
1	0	Jackson	-0.232	0.089	0.009	0.038	0.098	0.696
		Minneapolis Suburbs	-0.502	0.080	< 10 ⁻⁴	-0.033	0.052	0.527
		Washington County	0.114	0.077	0.136	0.086	0.050	0.085
		Age	-0.007	0.005	0.191	0.103	0.004	< 10 ⁻⁴
		Male	-0.030	0.054	0.578	0.562	0.036	< 10 ⁻⁴

r _G	r _H	Covariate	Diabete			Death		
			Est	Std. Err	p-value	Est	Std. Err	p-value
		White	-0.629	0.112	<10 ⁻⁴	-0.500	0.102	<10 ⁻⁴
		Body mass index	0.094	0.005	<10 ⁻⁴	-0.003	0.004	0.411
		Glucose	0.122	0.003	<10 ⁻⁴	0.006	0.002	0.001
		Systolic blood pressure	0.007	0.002	0.001	0.017	0.001	<10 ⁻⁴
		Diastolic blood pressure	-0.0005	0.004	0.893	-0.014	0.002	<10 ⁻⁴

Note: Forsyth County, NC, is the reference group for the field center variables.

Table 4
Regression analysis for diabetes in the ARIC study without adjustments for death

r_G	Covariate	Est	Std. Err	p-value
0	Jackson	-0.141	0.102	0.169
	Minneapolis Suburbs	-0.373	0.062	$<10^{-4}$
	Washington County	0.095	0.058	0.099
	Age	-0.010	0.004	0.009
	Male	-0.042	0.041	0.303
	White	-0.336	0.106	0.001
	Body mass index	0.067	0.003	$<10^{-4}$
	Glucose	0.090	0.002	$<10^{-4}$
	Systolic blood pressure	0.005	0.002	0.004
	Diastolic blood pressure	0.000	0.003	0.992
1	Jackson	-0.210	0.134	0.116
	Minneapolis Suburbs	-0.464	0.076	$<10^{-4}$
	Washington County	0.112	0.072	0.121
	Age	-0.013	0.005	0.009
	Male	-0.062	0.052	0.230
	White	-0.062	0.052	0.230
	Body mass index	0.089	0.005	$<10^{-4}$
	Glucose	0.114	0.003	$<10^{-4}$
	Systolic blood pressure	0.006	0.002	0.007
	Diastolic blood pressure	0.001	0.003	0.813

Note: Forsyth County, NC, is the reference group for the field center variables.