Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(7) 1624–1639

DOI: 10.1177/09622802211023975
journals.sagepub.com/home/smm
SAGE

Semiparametric estimation of the proportional rates model for recurrent events data with missing event category

Feng-Chang Lin[1], Jianwen Cai[1], Jason P Fine[1], Elisabeth P Dellon[2] and Charles R Esther[2]

Abstract

Proportional rates models are frequently used for the analysis of recurrent event data with multiple event categories. When some of the event categories are missing, a conventional approach is to either exclude the missing data for a complete-case analysis or employ a parametric model for the missing event type. It is well known that the complete-case analysis is inconsistent when the missingness depends on covariates, and the parametric approach may incur bias when the model is misspecified. In this paper, we aim to provide a more robust approach using a rate proportion method for the imputation of missing event types. We show that the log-odds of the event type can be written as a semiparametric generalized linear model, facilitating a theoretically justified estimation framework. Comprehensive simulation studies were conducted demonstrating the improved performance of the semiparametric method over parametric procedures. Multiple types of *Pseudomonas aeruginosa* infections of young cystic fibrosis patients were analyzed to demonstrate the feasibility of our proposed approach.

Keywords

Cystic fibrosis, generalized partially linear model, polynomial spline, rate proportion, weighted estimating equation

1 Introduction

Recurrent event data with multiple categories frequently arise in medical science and population health studies. Different causes of hospitalizations, multiple strains of bacteria infections, and various types of treatment failures all belong to such data type. Taking cystic fibrosis (CF) as an example, recurrent *Pseudomonas aeruginosa* (PA) infections are commonly observed in patients with CF. PA infection includes mucoid and nonmucoid strains. Without appropriate treatment, the recurrent infections with mucoid strains often become persistent and chronic, causing increased CF mortality and morbidity.[1–3] As another example, patients who received renal transplants may have different types of recurrent infections.[4] End-stage renal disease patients who received continuous ambulatory peritoneal dialysis may have multiple types of treatment failures that make the patient switch to other dialysis methods.[5]

When modeling this kind of recurrent event data, a proportional rates model that is conditional only on the current value of covariates is commonly used.[6] Let $N^*_{ij}(t)$ denote the number of recurrent events up to time t for

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA
[2]Department of Medicine, University of North Carolina at Chapel Hill, NC, USA

Corresponding author:
Feng-Chang Lin, Department of Biostatistics, University of North Carolina at Chapel Hill, CB 7420, Chapel Hill, NC 27599-7420, USA.
Email: flin@bios.unc.edu

subject i ($i = 1, \ldots, n$) and event category j ($j = 1, \ldots, J$). Let $Z_{ij}(t)$ denote the column vector of covariates which are possibly time-varying. A proportional rates model proposed by Cai and Schaubel[6] is defined by

$$E\{dN^*_{ij}(t)|Z_{ij}(t)\} = \exp\{\beta_0^{\mathrm{T}} Z_{ij}(t)\} d\mu_{0j}(t) \tag{1}$$

where β_0 is the regression coefficient and $d\mu_{0j}(t)$ is the baseline rate function for the jth type of the recurrent event. Although β_0 is not indexed by event category j, the model is flexible enough to accommodate covariate effects that are specific to the event type in each individual model. For example, if there are two types of recurrent events and one uses q_1- and q_2-column vector of covariates, $\tilde{Z}_{i1}$ and $\tilde{Z}_{i2}$, for the first and second type of recurrent events, respectively, one can define $Z_{i1} = (\tilde{Z}_{i1}^{\mathrm{T}}, 0_{q_2}^{\mathrm{T}})^{\mathrm{T}}$ and $Z_{i2} = (0_{q_1}^{\mathrm{T}}, \tilde{Z}_{i2}^{\mathrm{T}})^{\mathrm{T}}$ to specify the individual models, where 0_{q_j} is a q_j-column vector of zeros. Accordingly, one may define $\beta_0 = (\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}$, where $\beta_j = (\beta_{j1}, \ldots, \beta_{jq_j})^{\mathrm{T}}$ for $j = 1, 2$.

Let $Y_i(t) = I(C_i \geq t)$ indicate whether subject i with censoring time C_i is under observation at time t, where $t \in [0, \tau]$, and τ is the end of follow-up time. Let $N_{ij}(t) = Y_i(t) N^*_{ij}(t)$ denote the observed number of events up to time t. When the event category is fully observed, Cai and Schaubel[6] showed that the coefficient β_0 in (1) could be consistently estimated by the estimating equations

$$U^n(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^{\tau} \{Z_{ij}(t) - \bar{Z}_j(t;\beta)\}\, \mathrm{d}N_{ij}(t) = 0 \tag{2}$$

where $\bar{Z}_j(t;\beta) = S_j^{(1)}(t;\beta)/S_j^{(0)}(t;\beta)$ with

$$S_j^{(d)}(t;\beta) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_{ij}(t)^{\otimes d} \exp\{\beta^{\mathrm{T}} Z_{ij}(t)\}$$

for $d = 0$, 1, where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^{\mathrm{T}}$ for a column vector a.

However, when the event category is possibly missing, the estimating equation (2) is not feasible since the quantity $dN_{ij}(t)$ is not always observable. A naive approach, which uses completely observed data that include only events with known type, can be valid if the event category is missing completely at random, but may give biased results if the missingness depends on the covariates. Schaubel and Cai[5] suggested that one rewrite $dN_{ij}(t)$ as

$$dN_{ij}(t) = R_i(t) dN_{ij}(t) + \{1 - R_i(t)\} \delta_{ij}(t) dN_{i\cdot}(t) \tag{3}$$

where $R_i(t)$ indicates whether the event category is observed, $\delta_{ij}(t)$ indicates whether the event category is type j, and $dN_{i\cdot}(t) = \sum_{j=1}^{J} dN_{ij}(t)$ indicates the total number of events at time t. Note that $dN_{i\cdot}(t)$ equals 0 or 1 since they assume events with different types do not occur simultaneously. They further suggested that one replace the unknown quantity $\delta_{ij}(t)$ with a consistent estimator for $p_{ij}(t)$, where

$$p_{ij}(t) = E\{\delta_{ij}(t) | dN_{i\cdot}(t) = 1, Z_{ij}(t)\}$$

One may consider parametric, multinomial logit models for estimation of $p_{ij}(t)$.[5,7] However, the association between the covariates and δ_{ij} may not be correctly specified, which may lead to inconsistent estimation. This motivates us to develop a more robust method that weakens the impact of model misspecification.

In this paper, we extend the proportional rates method previously developed for estimation of $p_{ij}(t)$[8] using a semiparametric approach that exploits a special form of the rate proportion of event type j to the overall rate function. Interestingly, under the proportional rates model (1), the ratio of two rate proportions can be expressed as

$$\log\{p_{ij}(t)/p_{iJ}(t)\} = \beta_0^{\mathrm{T}} X_{ij}(t) + \eta_{0j}(t) \tag{4}$$

where $X_{ij}(t) = Z_{ij}(t) - Z_{iJ}(t)$ and $\eta_{0j}(t) = \log\{d\mu_{0j}(t)/d\mu_{0J}(t)\}$. In fact, model (4) can be viewed as a generalized partially linear model. Under certain regularity conditions, one can estimate β_0 and η_{0j} simultaneously via semiparametric regression techniques such as local polynomials,[9] generalized additive models,[10] and polynomial spline functions.[11]

When there is no missing data in the event type, one can estimate $p_{ij}(t)$ using all of the data based on model (4). With missing event types, however, one can only estimate $p^c_{ij}(t)$ using completely observed data, where

$$p^c_{ij}(t) = E\{\delta_{ij}(t)|dN_{i\cdot}(t) = 1, R_i(t) = 1, Z_{ij}(t)\}$$

Letting $\pi_{ij}(t|Z_{ij}) = E\{R_i(t)|dN_{ij}(t) = 1, Z_{ij}(t)\}$ denote the probability of non-missingness given that the event type j occurs, one can show

$$\log\{p^c_{ij}(t)/p^c_{iJ}(t)\} = \beta_0^{\mathrm{T}} X_{ij}(t) + \eta_{0j}(t) + \kappa_j(t|Z_{ij}) \tag{5}$$

where $\kappa_j(t|Z_{ij}) = \log\{\pi_{ij}(t|Z_{ij})/\pi_{iJ}(t|Z_{iJ})\}$ is the log-ratio of non-missingness for two event types. Since model (5) is time-varying, estimation is complicated. A common assumption in the previous literature is that $\pi_{ij}(t|Z_{ij})$ does not depend on j and $\kappa_j(t|Z_{ij}) = 0$ for each j.[5,7,8] This assumption corresponds to missing at random (MAR) assumption when the missingness does not depend on unobserved information.[12] In this paper, we adopt the same assumption and assume $\kappa_j(t|Z_{ij}) = 0$.

Theoretical challenge remains when the semiparametric estimator of $p_{ij}(t)$ is substituted for the unknown $\delta_{ij}(t)$ in the estimating equations for β. Specifically, it is not clear if one can still obtain a $n^{1/2}$ convergence rate in the estimation of β since the convergence rate of $p_{ij}(t)$ estimation is generally slower than $n^{1/2}$ with a semiparametric approach. We will show that, under mild regularity conditions, our estimator of β converges at a $n^{1/2}$ rate to a normal distribution with variance that may be consistently estimated using a simple plug-in formula.

The remaining sections are organized as follows. In Section 2, we exploit a cubic B-spline function for the estimation of $p_{ij}(t)$ and propose general estimating equations for the regression coefficient β_0 and baseline mean function $\mu_{0j}(t)$ in model (1). Consistency and large sample normality of the estimators are shown in Section 3. Finite-sample performances evaluated by comprehensive simulation experiments are studied in Section 4. A real-data analysis on multiple types of PA infections in the United States 2016 Cystic Fibrosis Foundation Patient Registry is presented in Section 5. Conclusions and discussions on future research are presented in Section 6.

2 Estimation method

Assume that $\eta_{0j}(t)$ can be approximated by a cubic B-spline function

$$\tilde{\eta}_{0j}(t;\xi_j) = \xi_{j0} + \sum_{k=1}^{m+3} \xi_{jk} b_k(t)$$

where $b_k(t)$ $(k = 1, \ldots, m+3)$ are basis functions, m is the number of interior knots, and $\xi_j = (\xi_{j0}, \ldots, \xi_{j(m+3)})^{\mathrm{T}}$ is a vector of spline coefficients. Let $\theta_0 = (\beta_0^{\mathrm{T}}, \xi_0^{\mathrm{T}})^{\mathrm{T}}$, where $\xi_0 = (\xi_1^{\mathrm{T}}, \ldots, \xi_{J-1}^{\mathrm{T}})^{\mathrm{T}}$. One can estimate θ_0 by maximizing an approximate log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \int_0^{\tau} \ell_i(t;\beta,\xi) R_i(t)\,\mathrm{d}N_{i\cdot}(t) \tag{6}$$

where

$$\ell_i(t;\beta,\xi) = \sum_{j=1}^{J} \delta_{ij}(t) m^c_{ij}(t) - \log\left[\sum_{j=1}^{J} \exp\{m^c_{ij}(t)\}\right]$$

with $m^c_{ij}(t;\theta) = \beta^{\mathrm{T}} X_{ij}(t) + \xi^{\mathrm{T}} B_j(t)$, where $B_j(t)$ is a $(m+4)\times(J-1)$ column vector with $b(t) = (1, b_1(t), \ldots, b_{m+3}(t))^{\mathrm{T}}$ in the jth block for $j = 1, \ldots, J-1$, and $B_J(t) = 0$ for all t. The number of interior knots m can be selected by Akaike information criteria (AIC) that minimizes $-2\ell(\theta)$ plus two times the number of parameters in θ. However, other approaches such as generalized cross-validation that approximates the leave-one-out cross-validation may also be considered.[13]

Letting $\tilde{\theta} = (\tilde{\beta}^{\mathrm{T}}, \tilde{\xi}^{\mathrm{T}})^{\mathrm{T}}$ denote the maximizer of (6), one can estimate $p_{ij}(t)$ by

$$p_{ij}(t;\tilde{\theta}) = \frac{\exp\{\tilde{\beta}^{\mathrm{T}} X_{ij}(t) + \tilde{\xi}^{\mathrm{T}} B_j(t)\}}{\sum_{\ell=1}^{J} \exp\{\tilde{\beta}^{\mathrm{T}} X_{i\ell}(t) + \tilde{\xi}^{\mathrm{T}} B_\ell(t)\}}$$

By solving the estimating equations

$$U^r(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^\tau \{Z_{ij}(t) - \bar{Z}_j(t;\beta)\} \mathrm{d}N^r_{ij}(t;\tilde{\theta}) = 0 \tag{7}$$

where $dN^r_{ij}(t;\theta) = R_i(t)dN_{ij}(t) + \{1 - R_i(t)\}p_{ij}(t;\theta)dN_{i\cdot}(t)$, one can obtain our proposed estimator $\hat{\beta}^r$ for β_0. With β replaced by $\hat{\beta}^r$ in the estimating equation

$$\sum_{i=1}^{n} \sum_{j=1}^{J} \int_0^\tau [dN_{ij}(t) - Y_i(t)\exp\{\beta^{\mathrm{T}} Z_{ij}(t)\} \mathrm{d}\mu_{0j}(t)] = 0$$

one can obtain an empirical estimator $\hat{\mu}^r_{0j}(t;\hat{\beta}^r, \tilde{\theta})$ for the baseline mean function $\mu_{0j}(t)$, where

$$\hat{\mu}^r_{0j}(t;\hat{\beta}^r, \tilde{\theta}) = n^{-1} \sum_{i=1}^{n} \int_0^t S_j^{(0)}(t;\hat{\beta}^r)^{-1} \, \mathrm{d}N^r_{ij}(t;\tilde{\theta}) \tag{8}$$

Note that, although β is denoted the same in the log-rate ratio model (4) and proportional rates model (1), the β in the model (4) may not be fully identifiable. The covariate with β in model (4) is $X_{ij}(t) = Z_{ij}(t) - Z_{iJ}(t)$, which is the difference between $Z_{ij}(t)$ and $Z_{iJ}(t)$. If a covariate, for example, age is included and has the common effect on the rate function for all event categories, then the corresponding X_{ij} equals 0, and consequently, the corresponding component in β is not identifiable. Another situation is when a covariate is included in rate model for all event categories, but the effects are different for different event category. In this situation, what is estimable in β in model (4) is the difference between the effects of that covariate for different categories and not the effects themselves. Taking $J = 2$ for example, one can write $Z_{i1} = (\tilde{Z}_i, 0)^{\mathrm{T}}$ and $Z_{i2} = (0, \tilde{Z}_i)^{\mathrm{T}}$, where $\tilde{Z}_i$ is the covariate in the rates models for both event categories, for example, gender, and $\beta = (\beta_1, \beta_2)^{\mathrm{T}}$, where β_j is the effect of $\tilde{Z}_i$ on rate function for event category j for $j = 1$, 2. The difference between Z_{i1} and Z_{i2} is $X_{i1}(t) = (\tilde{Z}_i, -\tilde{Z}_i)^{\mathrm{T}}$, and one can write $\beta^{\mathrm{T}} X_{i1}(t) = (\beta_1 - \beta_2)\tilde{Z}_i$. From this expression, we can see that only the contrast $\beta_1 - \beta_2$ can be estimated from model (4), not β_1 and β_2 individually. Including some same set of covariates in the rate models for some event categories is common in practice. Therefore, it is not feasible to use the log-likelihood function (6) to estimate β in general. However, even though β in model (4) is not identifiable, our proposed method can still work because $p_{ij}(t)$ can still be consistently estimated using function (6).

Also note that the proportional rates model (1) with a baseline rate function specific for each event type is quite general. One may restrict the model with the baseline rate function to be proportional to a reference category J, meaning $d\mu_{0j}(t) = \gamma_{0j} d\mu_0(t)$, where $d\mu_0(t)$ is the baseline rate function of the reference category. The model can be written as

$$E\{dN^*_{ij}(t)|Z_{ij}(t), \triangle_j\} = \exp\{\beta_0^{\mathrm{T}} Z_{ij}(t) + \gamma_0^{\mathrm{T}} \triangle_j\} \mathrm{d}\mu_0(t) \tag{9}$$

where $\triangle_j$ is a column vector of 1 in the jth element and 0 otherwise with $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0(J-1)})^{\mathrm{T}}$ as the corresponding coefficient. One can show that the formula (5) becomes

$$\log\{p^c_{ij}(t)/p^c_{iJ}(t)\} = \beta_0^{\mathrm{T}} X_{ij}(t) + \gamma_{0j} + \kappa_j(t|Z_{ij}) \tag{10}$$

and estimate β_0 and γ_{0j} using completely observed data via parametric estimating equations

$$\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau X_{ij}(t)\{\delta_{ij}(t)-p_{ij}^c(t;\beta,\gamma)\}R_i(t)\,\mathrm{d}N_{i\cdot}(t)=0 \tag{11}$$

where

$$p_{ij}^c(t;\beta,\gamma)=\frac{\exp\{\beta^{\mathrm{T}}X_{ij}(t)+\gamma_{0j}\}}{\sum_{\ell=1}^{J}\exp\{\beta^{\mathrm{T}}X_{i\ell}(t)+\gamma_{0\ell}\}}$$

assuming $\kappa_j(t|Z_{ij})=0$ for $j=1,\ldots,J$.

With $\hat{\beta}^c$ and $\hat{\gamma}^c$ solving equation (11), one can obtain a more efficient estimator for β_0 and γ_0 using all of the events by replacing the unknown quantity $\delta_{ij}(t)$ with $p_{ij}^c(t;\hat{\beta}^c,\hat{\gamma}^c)$. This yields the estimating equations

$$U^r(\beta,\gamma)=\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau\{W_{ij}(t)-\bar{W}(t;\beta,\gamma)\}\,\mathrm{d}N_{ij}^r(t;\hat{\beta}^c,\hat{\gamma}^c)=0 \tag{12}$$

where $W_{ij}(t)=(Z_{ij}^{\mathrm{T}}(t),\triangle_j^{\mathrm{T}})^{\mathrm{T}}$, $\bar{W}(t;\beta,\gamma)=S^{(1)}(t;\beta,\gamma)/S^{(0)}(t;\beta,\gamma)$ with

$$S^{(d)}(t;\beta,\gamma)=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}Y_i(t)W_{ij}(t)^{\otimes d}\exp\{\beta^{\mathrm{T}}Z_{ij}(t)+\gamma^{\mathrm{T}}\triangle_j\}$$

for $d=0$, 1, and

$$dN_{ij}^r(t;\beta,\gamma)=R_i(t)dN_{ij}(t)+\{1-R_i(t)\}p_{ij}(t;\beta,\gamma)dN_{i\cdot}(t)$$

By comparing models (5) and (10), one can conduct a statistical test for the proportionality of the baseline rate functions by testing if $\eta_{0j}(t)$ is constant for all t, i.e., testing the null hypothesis $H_0:\eta_{0j}(t)=\gamma_{0j}$ for $t\geq 0$. Using our approach, the null hypothesis is equivalent to $H_0:\xi_{j1}=\ldots=\xi_{j(m+3)}=0$ for each j, which can be tested via a Wald-type test procedure. Under a more restricted model (9), the fully parametric model (10) for $p_{ij}(t)$ is somewhat different from the one proposed by Schaubel and Cai.[5] The fully parametric model for $p_{ij}(t)$ in Schaubel and Cai[5] includes more covariates than model (10), such as time of event occurrence t and number of previous events $N_{i\cdot}(t-)$.

3 Asymptotic theory

Large sample properties of the semiparametric estimators $\hat{\beta}^r$ and $\hat{\mu}_{0j}^r$ using model (5) will be derived in this section. The developments are more challenging than those in Schaubel and Cai,[5] which covers only the more restrictive parametric model (10). We first state our notations. Let

$$p_{ij}^c(t;\theta)=\exp\{m_{ij}^c(t;\theta)\}/\sum_{\ell=1}^{J}\exp\{m_{i\ell}^c(t;\theta)\}$$

and let

$$\dot{p}_{ij}^c(t;\theta)=\partial p_{ij}^c(t;\theta)/\partial\theta=p_{ij}^c(t;\theta)\{\tilde{X}_{ij}(t)-\sum_{\ell=1}^{J}\tilde{X}_{i\ell}(t)p_{i\ell}^c(t;\theta)\}$$

where $\tilde{X}_{ij}(t) = (X_{ij}(t)^{\mathrm{T}}, B_j(t)^{\mathrm{T}})^{\mathrm{T}}$. The score function of $n^{-1}\ell(\theta)$ can be written as $\mathbb{U}(\theta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{U}_{ij}(\theta)$, where

$$\mathbb{U}_{ij}(\theta) = \int_0^\tau \tilde{X}_{ij}(t)\{\delta_{ij}(t) - p_{ij}^c(t;\theta)\}R_i(t)\,\mathrm{d}N_{i\cdot}(t)$$

and the negative Hessian matrix of $n^{-1}\ell(\theta)$ can be written as

$$\mathbb{H}(\theta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau \tilde{X}_{ij}(t)\dot{p}_{ij}^c(t;\theta)R_i(t)\,\mathrm{d}N_{i\cdot}(t)$$

Regularity conditions, especially for the number of interior knots, are outlined here. These conditions are required for the proof of the large sample properties of our estimators.

a. Variables $\{N_{ij}(\cdot), Y_{ij}(\cdot), Z_{ij}(\cdot)\}_{j=1}^{J}$ $(i = 1, \ldots, n)$ are independent and identically distributed.
b. The distribution of censoring time C_i satisfies $P(C_i \geq \tau) > 0$ for each i.
c. The sample path of the covariates satisfies $|Z_{ij\ell}(0)| + \int_0^\tau |dZ_{ij\ell}(t)| < c_Z < \infty$ for every ℓ, where $Z_{ij\ell}$ is the ℓth element of the covariate Z_{ij}.
d. The limiting matrices $\Omega(\beta_0)$ and $\mathbb{H}(\theta_0)$ are positive-definite.
e. The baseline rate functions $d\mu_{0j}(\cdot), j = 1, \ldots, J$ are bounded away from zero and infinity on $[0, \tau]$.
f. The second derivative of $\eta_{0j}(\cdot)$ exists and satisfies Lipschitz condition of order ϵ on $[0, \tau]$ for $j = 1, \ldots, J$ for some $\epsilon \in (0, 1]$.
g. The number of interior knots satisfies $n^{1/(4+2\epsilon)} < m < n^{1/4}$.

Note that Conditions (a)–(e) are regularity conditions for recurrent event processes, outlined in Cai and Schaubel.[6] The smoothness condition in (f) is similar to the condition (C1) in Wang et al.[11] and enables estimation of $\eta_{0j}(t)$ using spline functions, with Condition (g) describing the number of parameters used in the spline functions relative to the sample size.

According to Wang et al.,[11] the convergence rate of the estimator of β_0 in model (5) is $n^{1/2}$ under Conditions (c)–(g), while the convergence rate of the nonparametric estimator of $\eta_{0j}(t)$ is slower than $n^{1/2}$. This makes that the convergence rate of the estimator for $p_{ij}(t)$ is slower than $n^{1/2}$. However, we can show that the convergence rate of our estimator for the regression parameter β_0 is $n^{1/2}$. The following theorem describes the large sample theory of our estimator. The detailed proof is given in Appendix 1.

Theorem 1 *Under Conditions (a)–(g), the estimator $\hat{\beta}^r$ is a consistent estimator of β_0 and $n^{1/2}(\hat{\beta}^r - \beta_0)$ converges in distribution to a normal variable with mean 0 and variance Σ, which can be consistently estimated by $\hat{\Omega}(\hat{\beta}^r)^{-1}\hat{\Phi}(\hat{\beta}^r)\hat{\Omega}(\hat{\beta}^r)^{-1}$, where*

$$\hat{\Omega}(\beta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau \left\{S_j^{(2)}(t;\beta)/S_j^{(0)}(t;\beta) - \bar{Z}_j(t;\beta)^{\otimes 2}\right\}\mathrm{d}N_{ij}^r(t;\tilde{\theta})$$

and

$$\hat{\Phi}(\beta) = n^{-1}\sum_{i=1}^{n}\hat{\Psi}_i(\beta, \tilde{\theta})^{\otimes 2}$$

with

$$\hat{\Psi}_i(\beta, \theta) = \sum_{j=1}^{J}\int_0^\tau \{Z_{ij}(t) - \bar{Z}_j(t;\beta)\}d\hat{M}_{ij}^r(t;\beta,\theta) + \hat{\Gamma}(\beta,\theta)\mathbb{H}(\theta)^{-1}\mathbb{U}_{ij}(\theta),$$
$$d\hat{M}_{ij}^r(t;\beta,\theta) = dN_{ij}^r(t;\theta) - Y_i(t)\exp\{\beta^{\mathrm{T}}Z_{ij}(t)\}d\hat{\mu}_{0j}^r(t;\beta,\theta),$$

and

$$\hat{\Gamma}(\beta,\theta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau \{Z_{ij}(t) - \bar{Z}_j(t;\beta)\}\rho_{ij}(t;\theta)\tilde{X}_{ij}(t)^{\mathrm{T}}\{1 - R_i(t)\}dN_{i\cdot}(t)$$

where $\rho_{ij}(t;\theta) = p_{ij}(t;\theta)\{1 - p_{ij}(t;\theta)\}$.

The consistency of $\hat{\beta}^r$ can be proved via consistency of $\tilde{\theta}$ and conventional convex theories. Large sample normality can be established via an approximation to $n^{1/2}(\hat{\beta}^r - \beta_0)$ by a summation of independent and identical random vectors, as shown in Appendix 1. Note that the variation of $\hat{\beta}^r$ is larger than that for $\hat{\beta}^n$ which solves equation (2) assuming that there are no missing data, since the estimation for $p_{ij}(t)$ creates additional uncertainty when the event type is missing. This additional variation can be seen in the second term of $\hat{\Psi}_i(\beta,\theta)$. Empirical studies show that the efficiency loss compared to the estimator with no missing data may be rather small.

Let $A_j(t;\beta,\theta) = -\int_0^t \bar{Z}_j(s;\beta)d\hat{\mu}_{0j}(s;\beta,\theta)$, and

$$D_j(t;\beta,\theta) = n^{-1}\sum_{i=1}^{n}\int_0^t S_j^{(0)}(s;\beta)^{-1}\rho_{ij}(s;\theta)\tilde{X}_{ij}(s)\{1 - R_i(s)\}\,\mathrm{d}N_{i\cdot}(s)$$

The following theorem describes the limiting properties of the baseline mean function estimator $\hat{\mu}^r_{0j}(t;\hat{\beta}^r,\tilde{\theta})$ for $\mu_{0j}(t)$ in model (1).

Theorem 2 *Under the same conditions of Theorem 1, the baseline mean function estimator $\hat{\mu}^r_{0j}(t;\hat{\beta}^r,\tilde{\theta})$ is uniformly consistent for $\mu_{0j}(t)$, $t \in [0,\tau]$, and $n^{1/2}\{\hat{\mu}^r_{0j}(t;\hat{\beta}^r,\tilde{\theta}) - \mu_{0j}(t)\}$ converges weakly to a Gaussian process with mean 0 and covariance function $V_j(s,t)$, $s,t \in [0,\tau]$, which can be consistently estimated by*

$$\hat{V}^r_j(s,t) = n^{-1}\sum_{i=1}^{n}\hat{\phi}_{ij}(s;\hat{\beta}^r,\tilde{\theta})\hat{\phi}_{ij}(t;\hat{\beta}^r,\tilde{\theta}) \tag{13}$$

where

$$\begin{aligned}\hat{\phi}_{ij}(t;\beta,\theta) = {} & A_j(t;\beta,\theta)^{\mathrm{T}}\hat{\boldsymbol{\Omega}}(^{\beta)-1}\hat{\boldsymbol{\Psi}}_i(\beta,\theta)\\ & + D_j(t;\beta,\theta)\mathbb{H}(\theta)^{-1}\sum_{j=1}^{J}\mathbb{U}_{ij}(\theta) + \int_0^t S_j^{(0)}(s;\beta)^{-1}\,\mathrm{d}\hat{M}^r_{ij}(s;\beta,\theta)\end{aligned}$$

The proof begins by decomposing $\hat{\omega}_j(t) = \hat{\mu}^r_{0j}(t;\hat{\beta}^r,\tilde{\theta}) - \mu_{0j}(t)$ as $\hat{\omega}^{(1)}_j(t) + \hat{\omega}^{(2)}_j(t)$, where $\hat{\omega}^{(1)}_j(t) = \hat{\mu}^r_{0j}(t;\hat{\beta}^r,\tilde{\theta}) - \hat{\mu}^r_{0j}(t;\beta_0,\theta_0)$ and $\hat{\omega}^{(2)}_j(t) = \hat{\mu}^r_{0j}(t;\beta_0,\theta_0) - \mu_{0j}(t)$. The uniform consistency of $\hat{\mu}^r_{0j}(t;\hat{\beta}^r,\tilde{\theta})$ can be proved by showing that both $\sup_{t\in[0,\tau]}|\hat{\omega}^{(1)}_j(t)|$ and $\sup_{t\in[0,\tau]}|\hat{\omega}^{(2)}_j(t)|$ converge to 0. The uniform convergence of $\hat{\omega}^{(2)}_j(t)$ can be proved using a law of large numbers for empirical processes and uniform convergence of $\tilde{\eta}_{0j}$ to η_{0j}. The uniform convergence of $\hat{\omega}^{(1)}_j(t)$ involves some additional assumptions. The details of the proof are provided in Appendix 1. The proof of weak convergence, which establishes tightness and convergence to finite-dimensional distributions, follows the standard tools in Pollard[14] and van der Vaart and Wellner;[15] see Appendix 1 for details.

4 Simulation study

In this section, we demonstrate the feasibility of our proposed method via comprehensive simulations. We first examine a scenario when the proportional baseline rate model (9) holds; therefore, the general model (1) also holds. We then investigate a scenario when the baseline rate functions are nonproportional, in that model (1) holds, but not model (9). For subject i, two types of recurrent events were simulated from two intensity functions sharing the same latent variable G_i, which was sampled from Gamma$(1/\alpha,\alpha)$ with $E(G_i) = 1$ and $\mathrm{var}(G_i) = \alpha$. In the first scenario, the intensity functions are assumed $\lambda_{i1}(t) = G_i r_{01} t \exp(\beta_1 \tilde{Z}_i)$ and $\lambda_{i2}(t) = G_i r_{02} t \exp(\beta_2 \tilde{Z}_i)$ with constants r_{01} and r_{02}, while, in the second scenario, the intensity functions are $\lambda_{i1}(t) = G_i r_{01} \exp(\beta_1 \tilde{Z}_i)$ and $\lambda_{i2}(t) = G_i r_{02} h_0(t) \exp(\beta_2 \tilde{Z}_i)$, where $h_0(t) = \exp\{-\sin(t/3) - 3\cos(3t)\}$. We let $\alpha = 0$, 0.5, or 1 for different dependencies between two types of recurrent events, where $\alpha = 0$ indicates two types of recurrent events are independent. We set $r_{01} = 0.125$ and $r_{02} = 0.0625$ in the first scenario and $r_{01} = r_{02} = 0.125$ in the second scenario. We let $\beta_1 = 0$ or $\log(2)$ and $\beta_2 = 0$. The covariate $\tilde{Z}_i$ was randomly drawn from a Bernoulli distribution with probability 0.5. The censoring time was generated uniformly between 0 and 5 for each subject. In summary, there are 1.2–1.4 total number of events on average in these simulation scenarios, with a 1:2 or 1:3 ratio of type 1 events to type 2 events. The maximum number of events in a subject ranges from 6 to 11 events on average among the scenarios.

One can show that the rate functions can be expressed as $E\{dN^*_{ij}(t)|Z_{ij}(t)\} = \exp(\beta_0^{\mathrm{T}} Z_{ij}) d\mu_{0j}(t)$, where $\beta_0 = (\beta_1, \beta_2)^{\mathrm{T}}$, $Z_{ij} = (I(j=1)\tilde{Z}_i, I(j=2)\tilde{Z}_i)^{\mathrm{T}}$, $d\mu_{01}(t) = r_{01} t$ and $d\mu_{02}(t) = r_{02} t$ in the first scenario, and $d\mu_{01}(t) = r_{01}$ and $d\mu_{02}(t) = r_{02} h_0(t)$ in the second scenario. Note that the weighted estimating equations method in Schaubel and Cai[5] is unbiased in the first scenario if one uses $\tilde{Z}_i$ as the covariate in the logistic regression model for $p_{i1}(t)$. However, it is quite evident that the model is misspecified if one uses the same model in the second scenario.

We assumed that the probability of having a missing category was given by

$$1 - \pi_i(t) = [1 + \exp\{-\epsilon' z_i(t)\}]^{-1}$$

where $z_i(t) = (1, t, N_{i\cdot}(t-), \tilde{Z}_i)'$. We let $\epsilon = (\epsilon_0, \epsilon_t, \epsilon_n, \epsilon_z)'$, with $\epsilon_t = -0.15$, $\epsilon_n = 0.1$, and $\epsilon_z = 0, \log(2)$, where $\epsilon_z = 0$ indicated that the missingness depends on covariates and the missingness assumption is MAR. Various values of ϵ_0 were given to control the percentages of events with missing categories, denoted by $\mathcal{M}_p$. Here, we assume that the missingness does not depend on the event type, i.e., $\pi_{i1} = \pi_{i2} = \pi_i$ and $\kappa_1 = 0$.

Table 1 shows the simulation results for β_1 under 1,000 repetitions of sample size $n = 200$. We present the results of our proposed estimator $\hat{\beta}_1^r$ and the weighted estimating equations method $\hat{\beta}_1^w$, where $z_i(t)$ was used as a covariate in the logistic regression model for $p_{i1}(t)$ to derive $\hat{\beta}_1^w$. We also present the results of the estimator $\hat{\beta}_1^n$ assuming that there is no missing event category. This approach is generally not feasible in practice but provides the best possible results in the ideal situation. Note that our proposed estimator is obtained under a more general model (1). Later, we will show that our estimator endures little efficiency loss even when the underlying model has proportional baseline rates. We report the average of biases in our replicated estimates, empirical standard deviation σ_1, and the relative mean-squared error to our proposed method, denoted by $e_r^x = m^x/m^r$, where $m^x = (\hat{\beta}_1^x - \beta_1)^2 + (\sigma_1^x)^2$, $x = n, r, w$. The result shows that our estimator provides comparable estimates when the weighted estimating equations method correctly specifies the model in the first scenario. The efficiency loss is minimal, as the relative mean-squared errors are all close to 1. When the model is misspecified by the weighted estimating equations method in the second scenario, the estimator has a relatively larger variation, especially when more events with missing type are present in the data. Our estimator on the other hand provides a more robust approach with significant efficiency improvements. Overall, our proposed estimator does not lose much efficiency compared to the ideal solution, while outperforming the current existing parametric estimator.

Table 2 demonstrates that the distribution of our estimator can be well approximated by a normal distribution in finite samples. We show the results based on model (1) in the second scenario when $n = 200$ and 400. The results based on model (9) in the first scenario are similar; hence omitted here. As one can see, the average of our standard error estimates, denoted by $\bar{\sigma}_1$ and $\bar{\sigma}_2$, is close to the empirical standard deviation, σ_1 and σ_2, respectively, and the coverage rate $\mathcal{C}_p$ for β_1 based on the 95% confidence interval is close to the nominal level. Meanwhile, the size of the Wald-type test $\mathcal{C}_0$ for $\beta_2 = 0$ is close to the given significance level at 0.05.

As seen in Tables 1 and 2, our estimator is robust to the MAR assumption when $\epsilon_z \neq 0$. Both point and variance estimation are consistent. In fact, the efficiency is slightly better compared to the estimator when $\epsilon_z = 0$. We also set different values of κ_1 to examine the performance of our estimation method under the missingness not at random assumption. However, since the simulation results are similar, we do not report the results here.

5 CF registry data

CF is one of the most common life-shortening, autosomal recessive genetic disorders, affecting about 30,000 individuals in the United States.[16] It is caused by mutations in the gene encoding the CF transmembrane conductance regulator.[17] Chronic lung infection and associated inflammation lead to significant morbidity in CF, with respiratory failure the leading cause of mortality. PA, one of the major virulent pathogens in CF patients, is a

Table 1. Simulation results are reported based on scenario 1, where data were generated from model (9), and scenario 2, where data were generated from model (1).

					$\hat{\beta}_1 - \beta_1$			σ_1			Ratio	
Scenario	β_1	α	ϵ_z	$\mathcal{M}_p$ (%)	$\hat{\beta}_1^n$	$\hat{\beta}_1^r$	$\hat{\beta}_1^w$	$\hat{\beta}_1^n$	$\hat{\beta}_1^r$	$\hat{\beta}_1^w$	e_r^n	e_r^w
1	0.69	0.5	0	10	0.008	0.009	0.010	0.230	0.233	0.233	0.97	1.00
				20		0.011	0.011		0.237	0.236	0.95	0.99
				30		0.010	0.011		0.242	0.241	0.90	0.99
			0.69	10		0.010	0.010		0.234	0.234	0.97	1.00
				20		0.011	0.011		0.238	0.237	0.93	0.99
				30		0.011	0.011		0.241	0.240	0.91	0.99
		1.0	0	10	0.010	0.011	0.011	0.257	0.260	0.260	0.98	1.00
				20		0.012	0.012		0.260	0.260	0.98	1.00
				30		0.012	0.012		0.263	0.262	0.96	0.99
			0.69	10		0.012	0.012		0.260	0.259	0.98	1.00
				20		0.012	0.012		0.261	0.260	0.97	0.99
				30		0.013	0.013		0.261	0.260	0.97	0.99
2	0	0	0	20	–0.001	0.001	–0.003	0.249	0.267	0.279	0.87	1.09
				30		0.003	–0.007		0.280	0.300	0.80	1.15
				40		0.006	–0.005		0.294	0.322	0.72	1.20
			0.69	20		0.001	–0.005		0.265	0.280	0.88	1.11
				30		0.005	–0.005		0.278	0.297	0.81	1.14
				40		–0.006	–0.008		0.297	0.323	0.71	1.19
		0.5	0	20	–0.010	–0.015	–0.018	0.287	0.307	0.314	0.88	1.05
				30		–0.013	–0.020		0.321	0.339	0.80	1.12
				40		–0.013	–0.027		0.325	0.346	0.78	1.14
			0.69	20		–0.017	–0.021		0.307	0.316	0.87	1.06
				30		–0.015	–0.024		0.315	0.330	0.83	1.10
				40		–0.016	–0.022		0.333	0.357	0.75	1.15
3	0.69	0	0	20	0.002	0.008	0.007	0.222	0.238	0.245	0.87	1.06
				30		0.006	0.007		0.251	0.263	0.78	1.10
				40		0.005	0.006		0.266	0.281	0.70	1.12
			0.69	20		0.005	0.005		0.234	0.244	0.90	1.08
				30		0.007	0.008		0.247	0.258	0.81	1.09
				40		0.002	0.007		0.257	0.270	0.75	1.11
		0.5	0	20	–0.001	–0.003	–0.003	0.248	0.264	0.273	0.89	1.07
				30		–0.004	0.000		0.270	0.287	0.85	1.13
				40		–0.004	–0.001		0.283	0.302	0.77	1.14
			0.69	20		–0.003	–0.004		0.259	0.269	0.92	1.07
				30		–0.002	0.000		0.270	0.282	0.85	1.10
				40		–0.004	–0.004		0.274	0.287	0.82	1.10

well-known risk factor for CF lung disease progression and survival. Several baseline risk factors for PA acquisition were examined in Lai et al.[18] Meconium ileus, late CF diagnosis through signs and symptoms, severe CF genotypes, and female gender are associated with a higher risk of acquiring PA. However, most of the PA cases examined in the study were initial infections, which may be transient and less predictive of negative outcomes. On the other hand, mucoid PA, which is thought to develop after recurrent infections, is likely more critical to a patient's lung disease progression.[19] Therefore, regression modeling for different PA types, i.e., mucoid and nonmucoid, is important, since the baseline risk factors may differentially impact different infection types in various manners.

In this section, we extended the analysis in Lai et al.[18] to multiple event types using the United States 2016 CF Foundation Patient Registry (CFFPR), in which baseline characteristics, such as genotype, phenotype, and other prognosis factors, are recorded upon enrollment. The CFFPR documents the diagnosis and follow-up of 29,887 individuals with CF in the registry. We aim to model the nonmucoid and mucoid PA occurrence rates in association with three baseline risk factors, which include (1) gender, (2) genotype, categorized based on the most

Table 2. Simulation results of parameter estimations from our proposed method for model (1).

					$\hat{\beta}_1^r$				$\hat{\beta}_2^r$			
(β_1, β_2)	α	ϵ_z	$\mathcal{M}_p$ (%)	n	Bias	σ_1	$\bar{\sigma}_1$	$\mathcal{C}_p$	Bias	σ_2	$\bar{\sigma}_2$	$\mathcal{C}_0$
(0,0)	0	0	20	200	0.001	0.267	0.272	0.961	–0.006	0.149	0.151	0.035
				400	–0.002	0.185	0.191	0.959	–0.005	0.105	0.106	0.041
			30	200	0.003	0.280	0.287	0.961	–0.007	0.152	0.154	0.040
				400	–0.002	0.190	0.199	0.960	–0.005	0.107	0.108	0.045
			40	200	0.006	0.294	0.294	0.957	–0.007	0.153	0.155	0.038
				400	0.003	0.198	0.207	0.961	–0.006	0.109	0.109	0.049
		0.69	20	200	0.001	0.265	0.272	0.967	–0.006	0.149	0.151	0.041
				400	–0.003	0.184	0.191	0.964	–0.004	0.106	0.106	0.048
			30	200	0.005	0.278	0.282	0.961	–0.008	0.150	0.153	0.037
				400	–0.001	0.190	0.198	0.955	–0.005	0.107	0.108	0.046
			40	200	–0.006	0.297	0.295	0.950	–0.005	0.153	0.155	0.041
				400	–0.006	0.200	0.207	0.958	–0.004	0.109	0.110	0.050
	0.5	0	20	200	–0.015	0.307	0.297	0.948	0.000	0.195	0.188	0.067
				400	–0.004	0.212	0.207	0.944	–0.004	0.134	0.133	0.060
			30	200	–0.013	0.321	0.306	0.947	0.000	0.196	0.190	0.062
				400	–0.004	0.222	0.215	0.941	–0.004	0.135	0.134	0.057
			40	200	–0.013	0.325	0.317	0.951	–0.001	0.199	0.192	0.056
				400	–0.005	0.228	0.222	0.950	–0.004	0.136	0.136	0.057
		0.69	20	200	–0.017	0.307	0.297	0.946	0.000	0.194	0.188	0.062
				400	–0.005	0.213	0.207	0.948	–0.004	0.133	0.133	0.056
			30	200	–0.015	0.315	0.306	0.952	–0.001	0.197	0.190	0.064
				400	–0.003	0.220	0.214	0.945	–0.005	0.135	0.134	0.056
			40	200	–0.016	0.333	0.318	0.950	–0.001	0.198	0.192	0.064
				400	–0.010	0.234	0.223	0.950	–0.003	0.138	0.136	0.059
(0.69,0)	0	0	20	200	0.008	0.238	0.234	0.958	–0.006	0.158	0.152	0.060
				400	0.002	0.163	0.164	0.948	–0.001	0.110	0.107	0.063
			30	200	0.006	0.251	0.246	0.957	–0.005	0.164	0.157	0.065
				400	0.003	0.170	0.170	0.952	–0.001	0.111	0.109	0.053
			40	200	0.005	0.266	0.251	0.947	–0.004	0.167	0.158	0.068
				400	0.004	0.177	0.177	0.950	–0.002	0.113	0.111	0.059
		0.69	20	200	0.005	0.234	0.232	0.955	–0.006	0.159	0.152	0.066
				400	0.000	0.160	0.163	0.954	–0.001	0.110	0.107	0.061
			30	200	0.007	0.247	0.239	0.948	–0.007	0.163	0.155	0.064
				400	0.002	0.168	0.168	0.950	–0.002	0.112	0.109	0.056
			40	200	0.002	0.257	0.248	0.954	–0.005	0.169	0.159	0.067
				400	0.002	0.173	0.174	0.956	–0.001	0.115	0.112	0.060
	0.5	0	20	200	–0.003	0.264	0.260	0.952	0.004	0.201	0.189	0.071
				400	–0.004	0.178	0.183	0.956	0.001	0.136	0.133	0.058
			30	200	–0.004	0.270	0.268	0.950	0.004	0.203	0.191	0.068
				400	–0.006	0.183	0.189	0.961	0.001	0.138	0.135	0.055
			40	200	–0.004	0.283	0.277	0.946	0.004	0.204	0.193	0.064
				400	–0.005	0.191	0.195	0.960	0.001	0.140	0.137	0.049
		0.69	20	200	–0.003	0.259	0.259	0.952	0.004	0.201	0.189	0.070
				400	–0.005	0.176	0.182	0.964	0.001	0.137	0.134	0.060
			30	200	–0.002	0.270	0.265	0.955	0.003	0.201	0.191	0.068
				400	–0.005	0.180	0.187	0.962	0.000	0.138	0.135	0.066
			40	200	–0.004	0.274	0.272	0.952	0.003	0.205	0.193	0.063
				400	–0.004	0.184	0.191	0.962	–0.001	0.140	0.137	0.055

common mutation: F508del homozygous, F508del heterozygous, and neither or unknown, and (3) method of diagnosis, categorized in four groups described elsewhere:[18] newborn screening, meconium ileus, family history without symptom, and symptom and sign. Medication use for chronic PA infections and study site/center is included in the model for adjustment of possible confounding effects.

Table 3. Summary table for the complete-case analysis and rate proportion method.

	Complete-case			Rate proportion		
Covariates	exp(β)	95% CI	p-value	exp(β)	95% CI	p-value
	Nonmucoid PA infection					
Female	1.09	1.03, 1.15	0.001	1.08	1.01, 1.15	0.017
Genotype F508del			<0.001[a]			<0.001[a]
Homozygous	1.00	–	–	1.00	–	–
Heterozygous	0.89	0.83, 0.94	<0.001	0.89	0.84, 0.94	<0.001
Neither or unknown	0.89	0.77, 1.03	0.121	0.90	0.83, 0.97	0.008
Diagnostic method			0.079[a]			0.025[†]
Newborn screening	1.00	–	–	1.00	–	–
Meconium ileus	1.15	0.99, 1.34	0.069	1.10	1.01, 1.21	0.037
Family history	0.99	0.77, 1.27	0.946	0.94	0.80, 1.10	0.445
Symptom	1.13	0.93, 1.35	0.212	1.07	0.98, 1.16	0.127
Medication	1.85	1.64, 2.09	<0.001	1.88	1.68, 2.09	<0.001
	Mucoid PA infection					
Female	1.18	1.01, 1.38	0.038	1.16	1.08, 1.24	<0.001
Genotype F508del			0.298[a]			0.089[†]
Homozygous	1.00	–	–	1.00	–	–
Heterozygous	0.95	0.81, 1.12	0.547	0.95	0.86, 1.06	0.374
Neither or unknown	1.20	0.91, 1.59	0.199	1.21	1.00, 1.46	0.052
Diagnostic method			0.004[a]			<0.001[†]
Newborn screening	1.00	–	–	1.00	–	–
Meconium ileus	1.18	0.91, 1.54	0.212	1.13	0.99, 1.29	0.063
Family history	1.45	0.96, 2.21	0.079	1.38	1.02, 1.87	0.035
Symptom	1.59	1.17, 2.15	0.003	1.51	1.30, 1.75	<0.001
Medication	3.06	2.57, 3.66	<0.001	2.98	2.71, 3.28	<0.001
	Both PA infection					
Female	1.22	1.03, 1.44	0.018	1.21	1.12, 1.31	<0.001
Genotype F508del			0.287[a]			0.217[†]
Homozygous	1.00	–	–	1.00	–	–
Heterozygous	0.92	0.78, 1.09	0.341	0.93	0.81, 1.06	0.258
Neither or unknown	1.11	0.87, 1.42	0.388	1.13	0.90, 1.42	0.294
Diagnostic method			<0.001[a]			<0.001[a]
Newborn screening	1.00	–	–	1.00	–	–
Meconium ileus	1.62	1.24, 2.11	<0.001	1.54	1.32, 1.80	<0.001
Family history	1.91	1.11, 3.28	0.019	1.78	1.23, 2.58	0.002
Symptom	1.99	1.47, 2.70	<0.001	1.88	1.68, 2.11	<0.001
Medication	2.47	1.98, 3.08	<0.001	2.39	2.13, 2.68	<0.001

[a] Refer to the overall comparison among levels in genotype and diagnostic method.

In the 2016 registry, we identified 14,888 patients who were born after 1997 and had complete baseline risk data in 188 accredited CF centers. In summary, half of these patients were male, 47% were F508del homozygous, 39% were F508del heterozygous, 47% were diagnosed by newborn screening, 31% were diagnosed by emerging symptoms and signs, and 19% and 3% were diagnosed by meconium ileus and family history, respectively. In the follow-up visits, there were 27,288 nonmucoid and 6,323 mucoid PA infections, in addition to 5,445 culture positives for both nonmucoid and mucoid types at the same visit. Since our method assumes two kinds of events cannot occur simultaneously, we treated those visits with both infections as the third type of recurrent event. Meanwhile, there were 5,392 culture positives in PA but with unknown status, which is in a high frequency of missing event type (12% of the total events). Addressing the missing information is highly desirable.

Table 3 shows the estimation results by the complete-case analysis and our proposed rate proportion method, assuming that the missingness does not depend on the event category, i.e., $\kappa_1 = \kappa_2 = 0$. We used AIC to choose the number of interior knots in the B-spline function in model (5). Up to five interior knots were examined, the minimum value of AIC was achieved by using a cubic function without any interior knots. We report rate ratio, exp(β), and its 95% confidence interval (95% CI) with Wald-type test p-value with respect to a reference group.

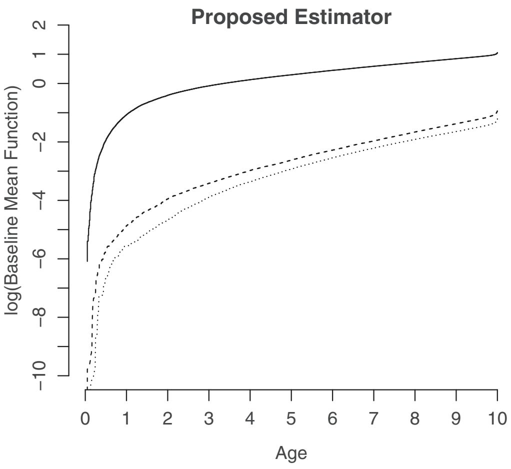

Figure 1. The baseline mean function estimation by our proposed method is shown in log-scale for nonmucoid (solid), mucoid (dash), and both (dot) PA infections.

We also report the p-value, indicated with a dagger, for the overall comparison among levels in genotype and diagnostic method.

As one can see, in the nonmucoid PA acquisition, both methods agree that female gender and heterozygous F508del genotype are significantly associated with the infection rate. While female has around 8% higher infection rate than male, patients with heterozygous F508del genotype have 11% lower infection rate than those with homozygous F508del genotype. However, the two methods have different test results in neither F508del nor unknown genotype and diagnosed by meconium ileus. While both methods have similar rate ratios, our proposed method, with more data points involved, has a narrower confidence interval and significant test result, which leads one to conclude that patients with a mild genotype (heterozygous F508del, neither F508del genotype, or unknown) have 10% lower nonmucoid PA infection rate than those with a more severe genotype (homozygous F508del), and patients diagnosed by with meconium ileus have 10% higher infection rate than those diagnosed by newborn screening. The finding is consistent with Lai et al.[18] and other scientific reports. The disparity between the complete-case analysis and our method can be considered as an evidence against the missingness completely at random assumption. In fact, based on a multiple logistic regression model for the likelihood of missingness, the probability of missing event type is significantly associated with age, gender, genotype, and frequency of previous events. The results demonstrate that missing event type is more likely to occur in younger patients, in female patients, in patients with mild genotypes, and in patients with more prior PA infections.

Two methods otherwise have similar results in the mucoid PA acquisition and in both the types of infection in the same culture. One difference is that patients diagnosed by family history without symptoms are statistically significant in the mucoid PA infection using our proposed method, but not significant using the complete-case analysis.

The baseline mean function estimation for the three types of infections by our proposed estimator is shown in Figure 1. The testing for the proportionality of the baseline rate functions based on the Wald-type test is significant with p-value <0.0001, suggesting that model (9) is not a good fit to our data. Hence, we only report the results for the model (1). We also assess the assumption of equal probability of missingness by assuming that the log-ratio of the missingness probability $\kappa_j(t|Z_{ij})$ in model (5) is possibly non-zero. Here, we implemented different values of κ_1 and κ_2, ranging from –1.5 to 1.5, to explore the sensitivity of the parameter estimation when the missingness is not at random. The result in the supplementary material shows that our estimation is quite robust to the violation of the MAR assumption, as the changes in the point and variance estimates of the regression coefficients are minimal, even when κ_1 and κ_2 are large.

6 Conclusion and discussion

It is worth noting that the same proportionality property exploited by our method was also discussed in an intensity-based recurrent event model[20] and in competing risk models with missing or uncertain cause of failure.[21–24] A semiparametric framework for the estimation of $p_{ij}(t)$ otherwise has never been explored. It is widely anticipated that the semiparametric estimation will be more robust if the underlying unknown function is indeed time-dependent, which is quite likely in practice with time to event data.

The estimation procedure in this paper is tailored for the proportional rates model. It may not be feasible for a nonproportional rates model since the ratio of the rate functions may not be log-linearly correlated with covariates. It would be of interest to develop a more general approach for different types of rate models. One possibility is to derive the probability of missingness and then inversely weigh the estimating equations for unbiased estimation. Along these lines, one may also utilize the nonparametric estimation for the rate function to construct a doubly robust estimator, providing additional protection against the model misspecification.

Acknowledgements

The authors would like to thank the Cystic Fibrosis Foundation for the use of CFFPR data to conduct this study. Additionally, the authors would like to thank the patients, care providers, and clinic coordinators at CF Centers throughout the United States for their contributions to the CFFPR. The authors also like to thank Dr Huichuan Lai for her thoughtful comments on an early draft of the paper.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002489, and by the National Cancer Institute (NCI) through Grant Award Number P01CA142538.

ORCID iD

Feng-Chang Lin https://orcid.org/0000-0002-2638-1775

References

1. Henry RL, Mellis CM and Petrovic L. Mucoid *Pseudomonas aeruginosa* is a marker of poor survival in cystic fibrosis. *Pediatr Pulmonol* 1992; **12**: 158–161.
2. Konstan MW, Morgan WJ, Butler SM et al. Risk factors for rate of decline in forced expiratory volume in one second in children and adolescents with cystic fibrosis. *J Pediatr* 2007; **151**: 134–139.
3. Li Z, Kosorok MR, Farrell PM et al. Longitudinal development of mucoid *Pseudomonas aeruginosa* infection and lung disease progression in children with cystic fibrosis. *JAMA* 2005; **293**: 581–588.
4. Chen X, Wang Q, Cai J et al. Semiparametric additive marginal regression models for multiple type recurrent events. *Lifetime Data Anal* 2012; **18**: 504–527.
5. Schaubel DE and Cai J. Rate/mean regression for multiple-sequence recurrent event data with missing event category. *Scand J Stat* 2006; **33**: 191–207.
6. Cai J and Schaubel D. Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Anal* 2004; **10**: 121–138.
7. Ye P, Zhao X, Sun L et al. A semiparametric additive rates model for multivariate recurrent events with missing event categories. *Comput Stat Data Anal* 2015; **89**: 39–50.
8. Lin FC, Cai J, Fine JP et al. Nonparametric estimation of the mean function for recurrent event data with missing event category. *Biometrika* 2013; **100**: 727–740.
9. Carroll RJ, Fan J, Gijbels I et al. Generalized partially linear single-index models. *J Am Stat Assoc* 1997; **92**: 477–489.
10. Hastie T and Tibshirani R. *Generalized additive models*. London, UK: Chapman & Hall, 1990.
11. Wang L, Liu X, Liang H et al. Estimation and variable selection for generalized additive partial linear models. *Ann Stat* 2011; **39**: 1827–1851.
12. Little RJA and Rubin DB. *Statistical analysis with missing data*. New York, NY: Wiley, 2002.
13. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York, NY: Springer, 2009.
14. Pollard D. *Empirical processes: theory and applications*. Hayward, CA: Institute of Mathematical Statistics, 1990.
15. van der Vaart AW and Wellner JA. *Weak convergence and empirical processes*. New York, NY: Springer, 1996.
16. Cystic Fibrosis Foundation. CF Foundation Patient Registry annual data report, 2017.
17. Riordan JR, Rommens JM, Kerem BS et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 1989; **245**: 1066–1073.

18. Lai HJ, Cheng Y, Cho H et al. Association between initial disease presentation, lung disease outcomes, and survival in patients with cystic fibrosis. *Am J Epidemiol* 2004; **159**: 537–546.
19. Folkesson A, Jelsbak L, Yang L et al. Adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis airway: an evolutionary perspective. *Nat Rev Microbiol* 2012; **10**: 841–851.
20. Chen BE and Cook RJ. The analysis of multivariate recurrent events with partially missing event types. *Stat Med* 2009; **24**: 671–691.
21. Craiu RV and Duchesne T. Inference based on the EM algorithm for the competing risks model with masked causes of failure. *Biometrika* 2004; **91**: 543–558.
22. Lu K and Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* 2001; **57**: 1191–1197.
23. Lu W and Liang Y. Analysis of competing risks data with missing cause of failure under additive hazards model. *Stat Sin* 2008; **18**: 219–234.
24. Racine-Poon AH and Hoel DG. Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics* 1984; **40**: 1151–1158.

Appendix I

A.1. Proof of theorem 1

By Taylor expansion of $U^r(\beta)$ around β_0, one can have $n^{1/2}(\hat{\beta}^r - \beta_0) = \hat{\Omega}(\bar{\beta})^{-1} n^{-1/2} U^r(\beta_0)$, where $\hat{\Omega}(\bar{\beta}) = -n^{-1}\partial U^r(\beta)/\partial\beta|_{\beta=\bar{\beta}}$ with $\bar{\beta}$ lying between $\hat{\beta}^r$ and β_0. Under Conditions (a)–(e), law of large numbers, consistency of $\hat{\beta}^r$ and $\tilde{\theta}$, and uniform convergence of $\tilde{\eta}_{0j}$, one can show that $\hat{\Omega}(\bar{\beta})$ converges to

$$\Omega(\beta_0) = \sum_{j=1}^{J} \int_0^{\tau} v_j(t;\beta_0) s_j^{(0)}(t;\beta_0)\, \mathrm{d}\mu_{0j}(t)$$

where for $d = 0$, 1, 2,

$$v_j(t;\beta) = s_j^{(2)}(t;\beta)/s_j^{(0)}(t;\beta) - \bar{z}_j(t;\beta)^{\otimes 2}$$

$$s_j^{(d)}(t;\beta) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E[Y_i(t) Z_{ij}(t)^{\otimes d} \exp\{\beta^{\mathrm{T}} Z_{ij}(t)\}]$$

and $\bar{z}_j(t;\beta) = s_j^{(1)}(t;\beta)/s_j^{(0)}(t;\beta)$. Let $U^r(\beta_0) = U_1^r(\beta_0) + U_2^r(\beta_0)$, where

$$U_1^r(\beta_0) = \sum_{i=1}^{n}\sum_{j=1}^{J} \int_0^{\tau} \{Z_{ij}(t) - \bar{Z}_j(t;\beta_0)\}\, \mathrm{d}M_{ij}^r(t;\beta_0)$$

with

$$dM_{ij}^r(t;\beta) = R_i(t) dN_{ij}(t) + \{1 - R_i(t)\} p_{ij}(t) dN_{i\cdot}(t) - Y_i(t) \exp\{\beta^{\mathrm{T}} Z_{ij}(t)\} d\mu_{0j}(t)$$

and

$$U_2^r(\beta_0) = \sum_{i=1}^{n}\sum_{j=1}^{J} \int_0^{\tau} \{Z_{ij}(t) - \bar{Z}_j(t;\beta_0)\}\{p_{ij}(t;\tilde{\theta}) - p_{ij}(t)\}\{1 - R_i(t)\}\, \mathrm{d}N_{i\cdot}(t)$$

Again, using Taylor expansion, one can get $p_{ij}(t;\tilde{\theta}) - p_{ij}(t) = \rho_{ij}(t;\bar{\theta})\{m_{ij}(t;\tilde{\theta}) - m_{ij}(t)\}$, where $m_{ij}(t) = \beta_0^{\mathrm{T}} X_{ij}(t) + \eta_{0j}(t)$ and $\bar{\theta} = (\bar{\beta}, \bar{\xi})$ satisfying $|m_{ij}(t;\bar{\theta}) - m_{ij}(t)| < |m_{ij}(t;\tilde{\theta}) - m_{ij}(t;\bar{\theta})|$ for every i and j. Rewriting

$m_{ij}(t;\tilde{\theta}) - m_{ij}(t) = (\tilde{\beta} - \beta_0)^{\mathrm{T}} X_{ij}(t) + (\tilde{\eta}_{0j} - \eta_{0j})(t)$, where $\tilde{\eta}_{0j}(t) = \tilde{\xi}^{\mathrm{T}} B_j(t)$, and letting $\rho_{ij}(t) = p_{ij}(t)\{1 - p_{ij}(t)\}$, one can show that both

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^{\tau}\{Z_{ij}(t) - \bar{z}_j(t;\beta_0)\}\rho_{ij}(t)(\tilde{\beta} - \beta_0)^{\mathrm{T}} X_{ij}(t)\{1 - R_i(t)\}\,\mathrm{d}N_{i\cdot}(t)$$

and

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^{\tau}\{Z_{ij}(t) - \bar{z}_j(t;\beta_0)\}\rho_{ij}(t)(\tilde{\eta}_{0j} - \eta_{0j})(t)\{1 - R_i(t)\}\,\mathrm{d}N_{i\cdot}(t)$$

are $o_p(n^{-1/2})$ when the number of interior knots follows Condition (g) since $\tilde{\beta} - \beta_0 = o_p(n^{-1/2})$ and $\int_0^{\tau}(\tilde{\eta}_{0j} - \eta_{0j})(t)dt = o_p(n^{-1/2})$, as derived by Wang et al.[11] Hence

$$n^{-1/2}U_2^r(\beta_0) = \Gamma(\beta_0,\theta_0)h(\theta_0)^{-1}n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{U}_{ij}(\theta_0) + o_p(1),$$

where $\Gamma(\beta_0,\theta_0) = \lim_{n\to\infty}\hat{\Gamma}(\beta_0,\theta_0)$ and $h(\theta_0) = \lim_{n\to\infty}\mathbb{H}(\theta_0)$.

Using conventional theories for empirical processes, one can also show that

$$n^{-1/2}U_1^r(\beta_0) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^{\tau}\{Z_{ij}(t) - \bar{z}_j(t;\beta_0)\}dM_{ij}^r(t;\beta_0) + o_p(1)$$

Accordingly, one has $n^{-1/2}U^r(\beta_0) = n^{-1/2}\sum_{i=1}^{n}\Psi_i(\beta_0,\theta_0) + o_p(1)$, where

$$\Psi_i(\beta,\theta) = \sum_{j=1}^{J}\int_0^{\tau}\{Z_{ij}(t) - \bar{z}_j(t;\beta)\}\,\mathrm{d}M_{ij}^r(t;\beta) + \Gamma(\beta,\theta)h(\theta)^{-1}\sum_{j=1}^{J}\mathbb{U}_{ij}(\theta)$$

The rest of the proof follows the central limit theorem.

A.2. Proof of theorem 2

First, we let $\hat{\omega}_j(t) = \hat{\mu}_{0j}^r(t;\hat{\beta}^r,\tilde{\theta}) - \mu_{0j}(t)$ and express $n^{1/2}\hat{\omega}_j(t)$ as $n^{1/2}\hat{\omega}_j(t) = n^{1/2}\hat{\omega}_j^{(1)}(t) + n^{1/2}\hat{\omega}_j^{(2)}(t)$, where $\hat{\omega}_j^{(1)}(t) = \hat{\mu}_{0j}^r(t;\hat{\beta}^r,\tilde{\theta}) - \hat{\mu}_{0j}^r(t;\beta_0,\theta_0)$ and $\hat{\omega}_j^{(2)}(t) = \hat{\mu}_{0j}^r(t;\beta_0,\theta_0) - \mu_{0j}(t)$. Furthermore, we let $\hat{\omega}_j^{(1)}(t) = \hat{\omega}_{j1}^{(1)}(t) + \hat{\omega}_{j2}^{(1)}(t)$, where

$$\hat{\omega}_{j1}^{(1)}(t) = n^{-1}\sum_{i=1}^{n}\int_0^{t}\{S_j^{(0)}(s;\hat{\beta}^r)^{-1} - S_j^{(0)}(s;\beta_0)^{-1}\}\,\mathrm{d}N_{ij}^r(s;\tilde{\theta})$$

and

$$\hat{\omega}_{j2}^{(1)}(t) = n^{-1}\sum_{i=1}^{n}\int_0^{t}S_j^{(0)}(s;\beta_0)^{-1}\{\mathrm{d}N_{ij}^r(s;\tilde{\theta}) - \mathrm{d}N_{ij}^r(s;\theta_0)\}$$

Recall that $A_j(t;\beta,\theta) = -\int_0^t\bar{Z}_j(s;\beta)d\hat{\mu}_{0j}(s;\beta,\theta)$. By Taylor expansion, weak law of large numbers, and consistency of $\tilde{\theta}$, one can claim that $\hat{\omega}_{j1}^{(1)}(t) = A_j(t;\beta_0,\theta_0)^{\mathrm{T}}(\hat{\beta}^r - \beta_0) + o_p(1)$ because $\partial\hat{\omega}_{j1}^{(1)}(t)/\partial\beta = A_j(t;\beta,\theta_0) + o_p(1)$, since $\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}dN_{ij}^r(s;\tilde{\theta}) = d\mu_{0j}(t)s_j^{(0)}(t;\beta_0)$ and $\lim_{n\to\infty}S_j^{(d)}(t;\beta_0) = s_j^{(d)}(t;\beta_0)$ for $d=0$, 1 uniformly in

$t \in [0, \tau]$. Under Conditions (a)–(g), one can show that $\sup_{t\in[0,\tau]} |\hat{\omega}_{j1}^{(1)}(t)|$ converges to 0 since $A_j(t; \beta_0, \theta_0)$ is bounded when $n \to \infty$ and $\hat{\beta}^r$ is consistent for β_0. Similarly, one can write $\hat{\omega}_{j2}^{(1)}(t) = n^{-1}\sum_{i=1}^{n} \int_0^t S_j^{(0)}(s; \beta_0)^{-1}\{p_{ij}(t; \tilde{\theta}) - p_{ij}(t; \theta_0)\}\{1 - R_i(s)\}dN_{i\cdot}(s)$. By writing $p_{ij}(t; \tilde{\theta}) - p_{ij}(t; \theta_0) = \rho_{ij}(t; \bar{\theta})\{m_{ij}(t; \tilde{\theta}) - m_{ij}(t; \theta_0)\}$, one can show that $\sup_{t\in[0,\tau]} |\hat{\omega}_{j2}^{(1)}(t)|$ converges to 0 since $\hat{\omega}_{j2}^{(1)}(t) = D_j(t; \beta_0, \theta_0)^{\mathrm{T}}(\tilde{\theta} - \theta) + o_p(1)$, consistency of $\tilde{\theta}$, and the fact that $D_j(t; \beta_0, \theta_0)$ is bounded when $n \to \infty$. Along with the uniform consistency of $\hat{\omega}_j^{(2)}(t)$, the uniform consistency of $\hat{\omega}_j(t)$ is proved.

By consistency of $\tilde{\theta}$ and $\hat{\mu}_{0j}(s; \beta_0, \tilde{\theta})$ in $s \in (0, t]$, one can claim that $A_j(t; \beta_0, \tilde{\theta})$ converges in probability to $a_j(t; \beta_0) = -\int_0^t \bar{z}_j(s; \beta_0)\, \mathrm{d}\mu_{0j}(s)$. To prove the large sample normality, one can show that $n^{1/2}\hat{\omega}_{j1}^{(1)}(t) = a_j(t; \beta_0)^{\mathrm{T}}\boldsymbol{\Omega}(\beta_0)^{-1}n^{-1/2}\sum_{i=1}^{n} \Psi_i(\beta_0, \theta_0) + o_p(1)$, and $n^{1/2}\hat{\omega}_{j2}^{(1)}(t) = d_j(t; \beta_0)h(\theta_0)^{-1}n^{-1/2}\sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{U}_{ij}(\theta_0) + o_p(1)$, where

$$d_j(t; \beta_0) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \int_0^t s_j^{(0)}(s; \beta_0)^{-1}\rho_{ij}(s)\tilde{X}_{ij}(s)\{1 - R_i(s)\}\mathrm{d}N_{i\cdot}(s)$$

Furthermore, one can write $n^{1/2}\hat{\omega}_j^{(2)}(t) = n^{-1/2}\sum_{i=1}^{n} \int_0^t s_j^{(0)}(s; \beta_0)^{-1}dM_{ij}^r(s; \beta_0, \theta_0) + o_p(1)$, where

$$dM_{ij}^r(t; \beta_0, \theta_0) = dN_{ij}^r(t; \theta) - Y_i(t)\exp\{\beta^T Z_{ij}(t)\}d\mu_{0j}(t)$$

Accordingly, the process $n^{1/2}\hat{\omega}_j(t)$ can be written as $n^{1/2}\hat{\omega}_j(t) = n^{1/2}\hat{\omega}_{j1}^{(1)}(t) + n^{1/2}\hat{\omega}_{j2}^{(1)}(t) + n^{1/2}\hat{\omega}_j^{(2)}(t)$, which can be expressed as $n^{-1/2}\sum_{i=1}^{n} \phi_{ij}(t; \beta_0, \theta_0) + o_p(1)$, where

$$\begin{aligned}\phi_{ij}(t; \beta_0, \theta_0) = {} & a_j(^{t;\beta_0)^{\mathrm{T}}}\boldsymbol{\Omega}(^{\beta_0)^{-1}}\Psi_i(\beta_0, \theta_0) \\ & + d_j(t; \beta_0)h(^{\theta_0)^{-1}}\sum_{j=1}^{J}\mathbb{U}_{ij}(\theta_0) + \int_0^t s_j^{(0)}(^{s;\beta_0)^{-1}}dM_{ij}^r(s; \beta_0, \theta_0)\end{aligned}$$

One can see that $n^{-1/2}\sum_{i=1}^{n} \phi_{ij}(t; \beta_0, \theta_0)$ is a normalized sum of independent and identically distributed random variables and, by the central limit theory, converges to a multivariate normal distribution with mean zero and covariance $V_j(s, t) = E\{\phi_{1j}(s; \beta_0, \theta_0)\phi_{1j}(t; \beta_0, \theta_0)\}$ given finitely many $s, t \in [0, \tau]$. Since $\phi_{1j}(t; \beta_0, \theta_0)$ is monotone in t, $n^{1/2}\hat{\omega}_j(t)$ is tight and hence converges weakly to a Gaussian process by the functional central limit theorem.