Published in final edited form as:
Biometrika. 2013 ; 100(3): . doi:10.1093/biomet/ast016.

Nonparametric estimation of the mean function for recurrent event data with missing event category

FENG-CHANG LIN,
Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A

JIANWEN CAI,
Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A

JASON P. FINE, and
Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A

HUICHUAN J. LAI
Department of Nutritional Sciences, University of Wisconsin, Madison, Wisconsin 53706, U.S.A

FENG-CHANG LIN: flin@bios.unc.edu; JIANWEN CAI: cai@bios.unc.edu; JASON P. FINE: jfine@bios.unc.edu; HUICHUAN J. LAI: hlai@wisc.edu

Summary

Recurrent event data frequently arise in longitudinal studies when study subjects possibly experience more than one event during the observation period. Often, such recurrent events can be categorized. However, part of the categorization may be missing due to technical difficulties. If the event types are missing completely at random, then a complete case analysis may provide consistent estimates of regression parameters in certain regression models, but estimates of the baseline event rates are generally biased. Previous work on nonparametric estimation of these rates has utilized parametric missingness models. In this paper, we develop fully nonparametric methods in which the missingness mechanism is completely unspecified. Consistency and asymptotic normality of the nonparametric estimators of the mean event functions accommodate nonparametric estimators of the event category probabilities, which converge more slowly than the parametric rate. Plug-in variance estimators are provided and perform well in simulation studies, where complete case estimators may exhibit large biases and parametric estimators generally have a larger mean squared error when the model is misspecified. The proposed methods are applied to data from a cystic fibrosis registry.

Some key words

Cystic fibrosis; Local polynomial regression; Nelson–Aalen estimation; *Pseudomonas aeruginosa* infection; Rate proportion

1. Introduction

Recurrent event data frequently occur in biomedical studies where subjects may suffer from repeated symptoms, infections or hospitalizations. Such data also arise in industrial manufacturing when tested units or equipment may experience multiple failures and repairs. Often, such recurrent events can be categorized. Taking cystic fibrosis for example, patients may experience repeated *Pseudomonas aeruginosa* infections in early childhood and later acquire other mutated types of infection, which also occur recurrently even after aggressive antibiotic use (Li et al., 2005). However, the identification of the event category may not be complete due to technical difficulties. As demonstrated in § 5, such missingness poses challenges for the analysis of the rates of particular event types.

A common approach for the analysis of recurrent events is based on a rate function. In contrast to an intensity function approach, which conditions on all previous information, a rate function approach conditions only on the current value of covariates (Pepe & Cai, 1993; Lin et al., 2000; Cook & Lawless, 2007; Cook et al., 2009). Complete case analyses that censor missing event types lead to underestimation of either the intensity or the rate function (Schaubel & Cai, 2006). Cai & Schaubel (2004) studied a proportional rate model for multiple recurrent event processes, with unbiased estimation of the regression parameters but not the baseline rate function obtained with missingness completely at random. Schaubel & Cai (2006) later proposed an estimation procedure that is valid under weaker missingness assumptions and yields unbiased estimates of the baseline rate function. Parametric models were used to estimate the missingness probabilities, which were then used as weights in the usual rate model estimating equation. Chen & Cook (2009) specified a parametric frailty model to characterize dependence amongst the events and employed maximum likelihood analysis, which requires correct specification of rate models for all event types, as well as of the frailty distribution. In this paper, we consider nonparametric estimation of the rate function without specifying parametric models for the missingness or imposing restrictions on the models for other event types.

To formalize the data set-up, suppose that there are n independent subjects with K recurrent event categories. Let $N^*_{ik}(t)$ denote the total number of category k events occurring before time t for subject i, such that $\mathrm{d}N^*_{ik}(s) \in \{0,1\}$ and $\mathrm{d}N^*_{ik}(s)\,\mathrm{d}N^*_{i\ell}(s)=0$ for $k \neq \ell$. The mean function $\mu_k(t)=E\{N^*_{ik}(t)\}$ is continuous with a smooth derivative $r_k(t) = \mathrm{d}\mu_k(t)/\mathrm{d}t$. Let C_i denote the censoring time for subject i. The observed number of events is given by $N_{ik}(t)=N^*_{ik}(t \wedge C_i)$, where $a \wedge b$ denotes the minimum of a and b. Assuming C_i is independent of N^*_{ik} for each i and k, we have $E\{N_{ik}(t) \mid Y_i(t)\} = Y_i(t)\mu_k(t)$ with $Y_i(t) = I(C_i \geq t)$ indicating whether subject i is at risk for any event type.

With event categories always being observed, a Nelson–Aalen-type estimator (Nelson, 1988), defined by

$$\hat{\mu}^n_k(t)=\sum_{i=1}^{n}\int_0^t Y.(s)^{-1}\mathrm{d}N_{ik}(s), \quad (1)$$

is consistent for $\mu_k(t)$ for each k, where $Y.(t)=\sum_{i=1}^{n} Y_i(t)$ denotes the total number of subjects who are at risk at time t. The variance of $\hat{\mu}^n_k(t)$ can be consistently estimated by

$$\hat{V}^n_k(t)=\sum_{i=1}^{n}\left[\int_0^t Y.(s)^{-1}\{\mathrm{d}N_{ik}(s)\}-Y_i(s)\,\mathrm{d}\hat{\mu}^n_k(s)\}\right]^2.$$

In the previous literature, this estimator was studied only for events of a single type (Andersen et al., 1993; Lawless & Nadeau, 1995; Cook et al., 1996; Chiang et al., 2005). With multiple event types, one may choose to explicitly model the dependence amongst the events, e.g., using a mixed Poisson process (Abu-Libdeh et al., 1990) or to construct marginal models that may be fitted separately (Cai & Schaubel, 2004). Intuitively, $\hat{\mu}^n_k$ should behave like estimators with a single type, since the estimator is calculated separately for each *k*.

The estimator (1) cannot be computed when event category information is missing. Naively censoring such events in a complete case analysis leads to underestimation. On the other hand, even with such missingness, the overall event process $\mathrm{d}N_{i\cdot}(t)=\sum_{k=1}^{K}\mathrm{d}N_{ik}(t)$ is observable. Using this information and information on events with known event types, one may estimate the probabilities of different event types conditionally on the observed data. These probabilities may be incorporated as weights in (1), yielding valid inferences. Schaubel & Cai (2006) employed a fully parametric logit model for the event category probabilities. When the model is misspecified, the resulting estimate for $\mu_k(t)$ could be biased. In this paper, we develop a fully nonparametric method for estimating $\mu_k(t)$ that is able to estimate the probability of an event being type *k* without any model assumption. The event category probabilities cannot be estimated at the usual parametric rate, which greatly complicates the analysis of the weighted version of (1). We show that the resulting estimator is root-*n* consistent and asymptotically normal, with variance which may be estimated using a simple plug-in formula.

2. Estimation methods

Let $\delta_i(t) \in \{1,\ldots, K\}$ denote the type of the event that occurs to subject *i* at time *t*, and let $\delta_{ik}(t) = I\{\delta_i(t) = k\}$ be an indicator function that indicates the category. Let $R_i(t) = 1$ when the event category is observed and $R_i(t) = 0$ otherwise. When some of the event categories are missing, a complete case analysis based on events with known event types, which is defined by

$$\widehat{\mu}_k^c(t)=\sum_{i=1}^{n}\int_0^t Y_{\cdot}(s)^{-1}R_i(s)\,\mathrm{d}N_{ik}(s),$$

will underestimate $\mu_k(t)$ even when the event category is missing completely at random.

Note that $\mathrm{d}N_{ik}(t) = \delta_{ik}(t)\,\mathrm{d}N_{i\cdot}(t)$, since $\mathrm{d}N_{ik}(t)\,\mathrm{d}N_{i\ell}(t) = 0$ for $k \neq \ell$. Thus, $\mathrm{d}N_{ik}(t) = R_i(t)\,\mathrm{d}N_{ik}(t) + \{1 - R_i(t)\}\,\delta_{ik}(t)\,\mathrm{d}N_{i\cdot}(t)$, and $\widehat{\mu}_k^n(t)$ in (1) can be written as

$$\widehat{\mu}_k^c(t)+\sum_{i=1}^{n}\int_0^t Y_{\cdot}(s)^{-1}\{1-R_i(s)\}\delta_{ik}(s)\,\mathrm{d}N_{i\cdot}(s). \quad (2)$$

Since $\delta_{ik}(t)$ is unobservable when $R_i(t) = 0$, the complete case estimator $\widehat{\mu}_k^c(t)$ underestimates the truth due to ignorance of the second part in (2). A prediction of $\delta_{ik}(t)$, based on observable data, could be inserted to estimate the unknown part and correct the underestimation of $\widehat{\mu}_k^c(t)$.

Assume that $\pi_i(t) = E\{R_i(t) \mid \mathrm{d}N_{ik}(t) = 1\}$ is the same for each *k*. One can show that

$$p_k(t)=E\{\delta_{ik}(t)|\mathrm{d}N_{i\cdot}(t)=1, R_i(t)=0, Y_i(t)\}=\frac{\{1-\pi_i(t)\}E\{\mathrm{d}N_{ik}(t)|Y_i(t)\}}{\sum_{\ell=1}^{K}\{1-\pi_i(t)\}E\{\mathrm{d}N_{i\ell}(t)|Y_i(t)\}},$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

for $k = 1,\dots, K$, which equals $r_k(t)/\sum_{\ell=1}^{K} r_\ell(t)$. Thus, if one can estimate $p_k(t)$ based on the rate functions $r_k(t)$, a consistent estimator may be derived by inserting in the estimated probabilities for the missing $\delta_{ik}(t)$ in (2). However, it is not clear how to estimate the rate function $r_k(t)$ when events with missing type are present in the data. Interestingly, without estimating $r_k(t)$ for each k, one may estimate $p_k(t)$, a rate proportion, by utilizing the events with known type, i.e., from a complete case analysis.

One can show that the limiting processes of $\hat{\mu}_k^c(t)$ and its derivatives, respectively, are

$$\mu_k^c(t)=\int_0^t y_1(s)^{-1}\pi_1^*(s)\,\mathrm{d}\mu_k(s), \quad r_k^c(t)=y_1(t)^{-1}\pi_1^*(t)r_k(t),$$

where $y_1(t)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\{Y_i(t)\}$ and $\pi_1^*(t)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\pi_i(t)E\{Y_i(t)\}$. One may utilize $r_k^c(t)$ to estimate the rate proportion $p_k(t)$, using the fact that

$$p_k(t)=\frac{r_k(t)}{\sum_{\ell=1}^{K} r_\ell(t)}=\frac{y_1(t)\pi_1^*(t)^{-1}r_k^c(t)}{y_1(t)\pi_1^*(t)^{-1}\sum_{\ell=1}^{K} r_\ell^c(t)}=\frac{r_k^c(t)}{\sum_{\ell=1}^{K} r_\ell^c(t)}.$$

That is, although the complete case estimator itself underestimates the true underlying rate function, it can otherwise consistently estimate the probability of an observed event being type k. We hereafter refer to this approach as the rate proportion method, since the probability is simply a proportion of the overall rate.

To estimate $p_k(t)$, we propose a nonparametric estimator for $\theta_k(t) = \log\{p_k(t)/p_K(t)\}$ via a local likelihood method and estimate $p_k(t)$ through $p_k(t)=\exp\{\theta_k(t)\}/\sum_{\ell=1}^{K}\exp\{\theta_\ell(t)\}$. For any time $t_0 \in [0, \tau]$, define the νth derivative of $\theta_k(t)$ as $\theta_k^{(\nu)}(t)=\partial^\nu\theta_k(t)/\partial t^\nu$. One may expand $\theta_k(t)$ as

$$\theta_k(t) \approx \sum_{\nu=0}^{q}\frac{1}{\nu!}\theta_k^{(\nu)}(t_0)\,(t-t_0)^\nu,$$

if t is in the neighbourhood of t_0, say, $t \in [t_0 - h, t_0 + h]$ with bandwidth h. Let $\beta_{\nu k}=(\nu!)^{-1}\theta_k^{(\nu)}(t_0)$, $\beta_k = (\beta_{0k}, \dots, \beta_{qk})^{\mathrm{T}}$, and $\tilde{\theta}_k(t,t_0;\beta_k)=\sum_{\nu=0}^{q}\beta_{\nu k}(t-t_0)^\nu$. The local log-likelihood for $\beta=(\beta_1^{\mathrm{T}}\dots,\beta_{K-1}^{\mathrm{T}})^{\mathrm{T}}$ is defined by

$$\ell(\beta)=\sum_{i=1}^{n}\int_0^\tau \mathscr{K}_h(u-t_0)\ell_i(u,t_0;\beta)\,R_i(u)\,\mathrm{d}N_{i\cdot}(u),$$

with

$$\ell_i(u,,t_0;\beta)=\sum_{k=1}^{K-1}\delta_{ik}(u)\tilde{\theta}_k(u,t_0;\beta_k)-\log\left[1+\sum_{k=1}^{K-1}\exp\{\tilde{\theta}_k(u,t_0;\beta_k)\}\right],$$

where $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ being a kernel function; τ is a constant that satisfies $\mathrm{pr}(C_i \geq \tau) > 0$ for each i. By theory of local polynomial modelling (Fan & Gijbels, 1996), we can approximate $\theta_k(t)$ by $\tilde{\theta}_k(t,t_0;\widehat{\beta}_k)=\sum_{\nu=0}^{q}\widehat{\beta}_{\nu k}(t-t_0)^{\nu}$, where $\hat{\beta}_k = (\hat{\beta}_{0k},\ldots,\hat{\beta}_{qk})^{\mathrm{T}}$ maximizes the local likelihood $\ell(\beta)$. Consequently, an estimator for $\theta_k(t_0)$ is simply the local intercept $\hat{\beta}_{0k}$, and by moving t_0 within $[0, \tau]$, we can obtain functional estimates for $\theta_k(t)$.

Our goal, however, is to replace $\delta_{ik}(t)$ in (2) with an estimate of $p_k(t)$ by

$$p_k(t;\widehat{\theta})=\frac{\exp(\widehat{\beta}_{0k})}{\sum_{\ell=1}^{K}\exp(\widehat{\beta}_{0\ell})},$$

where $\hat{\theta} = (\hat{\beta}_{01},\ldots,\hat{\beta}_{0(K-1)})^{\mathrm{T}}$ with $\hat{\beta}_{0k}$ $(k = 1,\ldots, K-1)$ being local likelihood estimates at t, and $\hat{\beta}_{0K} \equiv 0$. Our estimator of the mean function by the rate proportion method is

$$\widehat{\mu}_k^r(t;\widehat{\theta})=\widehat{\mu}_k^c(t)+\sum_{i=1}^{n}\int_0^t Y.(s)^{-1}\{1-R_i(s)\}p_k(s;\widehat{\theta})\,\mathrm{d}N_{i\cdot}(s), \quad (3)$$

with consistent variance estimator

$$\widehat{V}_k^r(t)=n^{-1}\sum_{i=1}^{n}\widehat{\varphi}_{ik}(t;\widehat{\theta})^2, \quad (4)$$

where $\varphi_{ik}(t;\hat{\theta})$ is defined in Theorem 1 in § 3 and $p_k(t)$ is estimated only when an event with unknown category occurred at t.

3. Asymptotic properties

Let $A(\upsilon)$ be a column vector that satisfies $\tilde{\theta}_k(u,t;\beta_k)=\beta_k^{\mathrm{T}}A(u-t)$ and $A(\upsilon)^{\otimes 2} = A(\upsilon)A(\upsilon)^{\mathrm{T}}$. Take $\tilde{p}_k(u,t;\beta)=\exp\{\tilde{\theta}_k(u,t;\beta_k)\}/[1+\sum_{\ell=1}^{K-1}\exp\{\tilde{\theta}_\ell(u,t;\beta_\ell)\}]$ for $k = 1,\ldots, K \ldots 1$. In addition, let $\{(K-1)\times(q+1)\}$-square matrix $\mathbb{H}$ denote blockdiag$\{H,\ldots, H\}$ with $H = \mathrm{diag}\{1, h, \ldots, h^q\}$, and take $\hat{\beta}^* = \mathbb{H}\hat{\beta}$ and $\beta_0^*=\mathbb{H}\beta_0$, where β_0 is the true value of β. Let $r.(u)=\sum_{k=1}^{K}r_k(u)$ and $\theta^{(\nu)}(u)=\{\theta_1^{(\nu)}(u),\ldots,\theta_{K-1}^{(\nu)}(u)\}^{\mathrm{T}}$.

We first provide the following lemma showing the consistency and large sample normality of the local likelihood estimator, which can be derived from a local polynomial method (Fan & Gijbels, 1996).

Lemma 1

Assume that the regularity conditions in the Appendix hold. Given $t_0 \in [0, \tau]$, we have

$$(nh)^{1/2}\{\widehat{\beta}^{*}-\beta_0^{*}-b(t_0)\}\to N\{0,\mathbb{A}(t_0)^{-1}\mathbb{B}(t_0)\mathbb{A}(t_0)^{-1}\}$$

in distribution, where $b(t_0)=\mathbb{A}(t_0)^{-1}\{\bar{\ell}_1(\beta_0^{*})^{\mathrm{T}},\ldots,\bar{\ell}_{K-1}(\beta_0^{*})^{\mathrm{T}}\}^{\mathrm{T}}$, with

$$\bar{\ell}_k(\beta_0^{*})=f(t_0)h^{q+1}\frac{\theta^{(q+1)}(t_0)^{\mathrm{T}}}{(q+1)!}\Omega_k(t_0)\int \upsilon^{q+1}A(\upsilon)\mathscr{K}(\upsilon)\,\mathrm{d}\upsilon\{1+o(1)\},$$

$f(t_0)=\pi_1^{*}(t_0)r.(t_0)$, $\Omega_k(t_0)$ is a $(K-1)$-column vector with $\rho_k(t_0)=p_k(t_0)\{1-p_k(t_0)\}$ in the kth element and $\rho_{k\ell}(t_0)=-p_k(t_0)\,p_\ell(t_0)$ in the ℓth element, for $\ell\neq k$; $\mathbb{A}(t_0)$ consists of diagonal block elements $\mathbb{A}_{kk}(k=1,\ldots,K\ldots 1)$, and off-diagonal block elements $\mathbb{A}_{k\ell}=\mathbb{A}_{\ell k}$, $k\neq\ell$, where $\mathbb{A}_{kk}=\rho_k(t_0)\,f(t_0)\int A(\upsilon)^{\otimes 2}\,\mathcal{K}(\upsilon)\,\mathrm{d}\upsilon$ and $\mathbb{A}_{k\ell}=\rho_{k\ell}(t_0)\,f(t_0)\int A(\upsilon)^{\otimes 2}\,\mathcal{K}(\upsilon)\,\mathrm{d}\upsilon$; $\mathbb{B}(t_0)$, the limiting variance matrix of the score function, consists of block elements $\mathbb{B}_{kk}=\rho_k(t_0)\,f(t_0)\int A(\upsilon)^{\otimes 2}\,\mathcal{K}(\upsilon)^2\,\mathrm{d}\upsilon$, and $\mathbb{B}_{k\ell}=\mathbb{B}_{\ell k}=\rho_{k\ell}(t_0)\,f(t_0)\int A(\upsilon)^{\otimes 2}\,\mathcal{K}(\upsilon)^2\,\mathrm{d}\upsilon$, for $k\neq\ell$.

In the special case with $q=1$ and $K=2$, Lemma 1 can be simplified to the following corollary.

Corollary 1

Under the conditions of Lemma 1, we have

$$(nh)^{1/2}\{\widehat{\beta}^{*}-\beta_0^{*}-b(t_0)\}\to N\{0,Q(t_0)\}$$

in distribution, where the bias $b(t_0)=(h^2/2)\mu_2\{\theta_1^{(2)}(t_0),0\}^{\mathrm{T}}+o(h^2)$ and the variance $Q(t_0)=\rho_1(t_0)^{-1}f(t_0)^{-1}\mathrm{diag}\{\upsilon_0,\mu_2^{-2}\upsilon_2\}$ with $\mu_2=\int \upsilon^2\,\mathcal{K}(\upsilon)\,\mathrm{d}\upsilon$, $\nu_0=\int \mathcal{K}(\upsilon)^2\,\mathrm{d}\upsilon$, and $\nu_2=\int \upsilon^2\,\mathcal{K}(\upsilon)^2\,\mathrm{d}\upsilon$. Furthermore,

$$(nh)^{1/2}\left\{\widehat{\theta}(t_0)-\theta_1(t_0)-\frac{1}{2}h^2\mu_2\theta_1^{(2)}(t_0)+o(h^2)\right\}\to N\{0,\sigma^2(t_0)\}$$

in distribution, where $\sigma^2(t_0)=\nu_0\rho_1(t_0)^{-1}f(t_0)^{-1}$.

When $q=1$ and $K=2$, the theoretical optimal bandwidth for estimating $\theta_1(\cdot)$ can be derived by minimizing the asymptotic integrated mean squared error $\int\{b(s)^2+\sigma^2(s)/(nh)\}\omega(s)\,\mathrm{d}s$ with some weighting function ω. One can show that

$$h_{\mathrm{opt}}=\{\int\sigma^2(s)w(s)\,\mathrm{d}s\}^{1/5}\left\{\mu_2^2\int\theta_1^{(2)}(s)^2w(s)\,\mathrm{d}s\right\}^{1/5}n^{-1/5}.$$

For arbitrary $K\geq 2$, one can show that the optimal choice of the bandwidth for $\theta_k(\cdot)$ is of order $n^{-1/(2q+3)}$ for $q\geq 0$. This is a critical result for the proof of the root-n weak convergence rate for $\widehat{\mu}_k^{r}$, due to the slower convergence rate of the local polynomial estimator $\widehat{\theta}$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Large sample properties of $\hat{\mu}_k^r$ are summarized in the following theorem, whose proof is given in the Appendix.

Theorem 1

Under the conditions of Lemma 1, the rate proportion estimator $\hat{\mu}_k^r(t;\hat{\theta})$ is uniformly consistent for $\mu_k(t)$ in $t \in [0, \tau]$, and $n^{1/2}\{\hat{\mu}_k^r(t;\hat{\theta})-\mu_k(t)\}$ converges weakly to a Gaussian process with mean zero and covariance function $V_k(s, t)$, $s, t \in [0, \tau]$, which can be consistently estimated by

$$\hat{V}_k^r(s,t)=\sum_{i=1}^{n}\hat{\varphi}_{ik}(s;\hat{\theta})\hat{\varphi}_{ik}(t;\hat{\theta}), \quad (5)$$

where

$$\hat{\varphi}_{ik}(t;\theta)=\int_0^t Y.(s)^{-1}\hat{\Omega}_k(s)^{\mathrm{T}}e_\nu^{\mathrm{T}}\hat{b}(s)\{1-R_i(s)\}\,\mathrm{d}N_{i\cdot}(s)+\int_0^t Y.(s)^{-1}\mathrm{d}\widehat{M}_{ik}^r(s;\hat{\theta}),$$

with $\hat{\Omega}_k(s)$ being a consistent estimate of $\Omega_k(s)$ obtained by replacing $p_k(s)$ with $p_k(s;\hat{\theta})$ for $k=1,\ldots, K-1$, $e_\nu=(e_1^{\mathrm{T}},\ldots,e_1^{\mathrm{T}})^{\mathrm{T}}$ with $(q + 1)$-column vectors $e_1 = (1, 0,\ldots, 0)^{\mathrm{T}}$, $\hat{b}(s)$ being an estimate of the bias term $b(s)$, and

$$\mathrm{d}\widehat{M}_{ik}^r(s;\theta)=R_i(s)\,\mathrm{d}N_{ik}(s)+\{1-R_i(s)\}\hat{p}_k(s;\theta)\,\mathrm{d}N_{i\cdot}(s)-Y_i(s)\,\mathrm{d}\hat{\mu}_k^r(s;\theta).$$

The summation of the first term in $\hat{\varphi}_{ik}(t;\theta)$ will be dominated by the summation of the second term. Hence the naive variance estimator for $\hat{\mu}_k^r(t)$, defined by

$$n^{-1}\sum_{i=1}^{n}\left\{\int_0^t Y.(s)^{-1}\mathrm{d}\widehat{M}_{ik}^r(s;\hat{\theta})\right\}^2,$$

is applicable when the sample size is large, without considering the variation contributed by the local likelihood estimates. That is, the limiting variance equals that from an estimator in which the event category probabilities are known. This differs from the case where parametric missingness models are fitted (Schaubel & Cai, 2006), where the resulting variance estimators depend on the variability in the parametric model estimates.

Observe, however, that the weak convergence rate of the two summation terms can be very close, e.g., $O(n^{-3/5})$ versus $O(n^{-1/2})$, when applying the local linear model. The naive variance estimator will likely underestimate the true variance when the sample size is relatively small, while the proposed variance estimator in (5) incorporates the variability of the local polynomial estimate. Specifically, one can estimate the bias term $b(s)$ by using a higher order polynomial. For example, in the special case with $q = 1$ and $K = 2$, the bias term depends on the second derivative of $\theta_1(t)$, which can be estimated by $2\hat{\beta}_{21}$ in a local cubic regression for $\theta_1(t)$. In short, we denote $\hat{V}_k^r(t)=\hat{V}_k^r(t,t)$, as in (4).

4. Simulation studies

In this section, simulation experiments are presented to demonstrate finite sample properties of our proposed estimation procedures. Three methods were evaluated. In the analysis of event category always being observed, we include every event in the estimation to serve as a reference for comparison. This kind of analysis is not feasible in practice with missing category data. Another method is the weighted estimating equations method (Schaubel & Cai, 2006) with a parametric logit model for the probability of a target category. A biased estimate may be anticipated when the true model is misspecified by the parametric model. Our proposed method, however, aims to provide consistent and robust estimates.

We consider three scenarios. In the first and second scenarios, we considered two types of recurrent events in 200 subjects. Let $\lambda_1(t) = 1$, $\lambda_2(t) = t$, and $\lambda_3(t) = t^2/3$. We first generated event processes with intensity functions $Gr_{01}\lambda_1(t)$ and $Gr_{02}\lambda_2(t)$, where the shared random variable G was sampled from a Gamma$(1/\alpha, \alpha)$ with $E(G) = 1$ and var$(G) = \alpha$. The mean functions we aim to estimate, therefore, are $\mu_1(t) = r_{01}t$ and $\mu_2(t) = r_{02}t^2/2$. In this setting the parametric logistic model in the weighted estimating equations method may correctly specify the model for $p_k(t)$ if one uses $\log(t)$ as a covariate since $\log\{p_1(t)/ p_2(t)\} = \log(r_{01}/ r_{02}) - \log(t)$. However, in a second scenario, if the second process is generated by an intensity function $Gr_{02}\{\lambda_1(t) + \lambda_2(t) + \lambda_3(t)\}$ with a mean function $\mu_2(t) = r_{02}(t + t^2/2 + t^3/9)$, the parametric model may be off the truth if one uses t as a covariate, especially when t is large. In the third scenario, we consider three types of recurrent events when $n = 50$ or 200 with intensity functions $Gr_{01}\lambda_1(t)$, $Gr_{02}\{\lambda_1(t) + \lambda_2(t)\}$, and $Gr_{02}\lambda_3(t)$, where $G = \log(W)/ \exp(0{\cdot}5)$ with W generated from a standard normal distribution.

The probability of having a missing category when an event occurred is

$$1-\pi_i(t)=[1+\exp\{-z_i(t)^{\mathrm{T}}\kappa\}]^{-1}, \quad (6)$$

where $z_i(t) = \{1, t, N_{i\cdot}(t\cdot), Z_i\}^{\mathrm{T}}$ with $N_{i\cdot}(t\cdot)$ counting the total number of events before t; $Z_i = 1$ if i is odd, and 0 otherwise. In the simulation we set $\kappa = (\kappa_0, \kappa_t, \kappa_n, \kappa_z)^{\mathrm{T}}$, with $\kappa_t = -0{\cdot}1$, $\kappa_n = 0{\cdot}05$, and $\kappa_z = 0$ or log(8), in which $\kappa_z \neq 0$ indicated missing due to covariates or missing at random in Little & Rubin (2002). Various values of κ_0 were set to create different amount of events with missing category in order to systematically explore the effects of missingness, for which estimators would have more variation when events with missing category occurred more often. The simulation results shown in Tables 1 and 2 support this.

We assumed $r_{01} = 0{\cdot}75$ or $1{\cdot}25$, $r_{02} = 0{\cdot}625$, and the Gamma parameter $\alpha = 0{\cdot}5$ or 1 in the first two scenarios, where a larger α represents higher dependence between event processes. On average, we observed about 4 events per subject when $\mu_2(t) = r_{02}t^2/2$ and about 7 events when $\mu_2(t) = r_{02}(t + t^2/2 + t^3/9)$. In the third scenario, we assumed $r_{01} = 0{\cdot}5$ and $r_{02} = 0{\cdot}625$, which also results in about 7 events per subject. Censoring times were independently generated by a uniform distribution between 0 and 5. All of our local likelihood estimation was implemented using the Epanechnikov kernel $\kappa(x) = 0{\cdot}75(1 - x^2)$, $|x| < 1$, and a local linear model, i.e., $q = 1$. When $K = 2$, a nearest-neighbour method was used to calculate the varying bandwidth and AIC (Akaike, 1974) was used as a bandwidth selection criteria. These procedures can be implemented using an R (R Development Core Team, 2013) package locfit (Loader, 2010). When $K = 3$, a fixed bandwidth proportional to $n^{-1/5}$ was applied. While $\log(t)$ was used in the first scenario for the correct model specification in the weighted estimating equations method, covariates $z_i(t)$ in (6) were used in the other two scenarios for the purpose of model misspecification.

We first show graphic results for $\mu_1(t)$ over the observation period with different combinations of α and κ_z in Fig. 1 when $\mu_2(t) = r_{02}t^2/2$. In these figures, the solid lines correspond to the true $\mu_1(t)$ and grey areas represent $\mu_1(t) \pm 1\cdot 96 \times \tilde{V}_1^r(t)^{1/2}$, where $\tilde{V}_1^r(t)$ is the empirical variance of the replicated estimates $\hat{\mu}_1^r(t)$; dotted lines show the average of the replicated $\hat{\mu}_1^r(t)$ and its $\pm 1\cdot 96 \times \overline{V}_1^r(t)^{1/2}$ pointwise confidence limits, where $\overline{V}_1^r(t)$ is the average of the replicated variance estimates $\hat{V}_1^r(t)$; dashed lines show the average of the replicated $\hat{\mu}_1^c(t)$ based on the complete case analysis. Overall, the estimation by the complete case analysis performs worse as the follow-up time t increases, due to more events with missing category at the later part of the observation period. On the contrary, our proposed estimator based on the rate proportion method is approximately unbiased. Also, the upper and lower dotted lines cover the grey area. This means that the point estimator $\hat{\mu}_1^r(t)$ is approximately unbiased and that the variance estimator $\hat{V}_1^r(t)$ approximates the asymptotic variance well.

Table 1 shows the simulation results for $\mu_1(t) = r_{01}t$ at $t = 3$ when $r_{01} = 0\cdot 75$ in the first two scenarios using $\hat{\mu}_1^n(t)$ in (1), the rate proportion method $\hat{\mu}_1^r(t)$ in (3), and the weighted estimating equations method $\hat{\mu}_1^w(t)$. We report the bias of the estimation, defined by the average of the replicated estimates minus the true value, the empirical standard deviation $\tilde{V}_1^{1/2}$, defined by the sample standard deviation of the replicated estimates, the average of the replicated standard deviation estimates $\overline{V}_1^{1/2}$, empirical coverage probability at a 0·95 nominal level, denoted by c_p, and the relative mean squared error to the rate proportion method, denoted by $e_r^x = m^x/m^r$, where $m^x = (\text{bias}^x)^2 + \tilde{V}_1^x\ (x = n, w)$ and m^r is defined similarly. The empirical percentage of recurrent events with missing category is denoted by $\mathcal{M}_p$. When $\mu_2(t) = r_{02}t^2/2$ and the weighted estimating equations method correctly specifies the model, all of the three estimators have bias close to 0 but $\hat{\mu}_1^r(t)$ has slightly larger empirical variance that results in a larger mean squared error. However, the relative error is rather moderate to $\hat{\mu}_1^n(t)$ and minimal to $\hat{\mu}_1^w(t)$. Hence our nonparametric estimator is very competitive with the current existing parametric method even when the parametric method correctly specifies the model. When $\mu_2(t) = r_{02}(t + t^2/2 + t^3/9)$ and the model was misspecified by the parametric method, only $\hat{\mu}_1^n(t)$ and $\hat{\mu}_1^r(t)$ are consistent. The estimator $\hat{\mu}_1^w(t)$ is generally biased and has larger empirical variance than $\hat{\mu}_1^r(t)$, resulting in a high ratio of mean squared errors. Overall, the rate proportion method has comparable variation to the analysis when the event category is always observed, has variance estimation close to the empirical variance that results in good empirical coverage, and has substantially better mean squared error when the true model is misspecified by the weighted estimating equations method. Similar results can be seen in Table 2, where the relative mean squared error is much greater in a later time when events with missing types occur more often. Interestingly, the empirical variance $\tilde{V}_1$ changes only slightly in both the rate proportion method and the weighted estimating equations method when the missingness depends on the covariate, so both estimators seem to be robust to the mechanics of missingness. However, when the data have more events with missing category, both estimators have larger variation but the rate proportion method performs better than the misspecified weighted estimating equations method.

5. Cystic fibrosis registry data

Cystic fibrosis is the most common life-shortening genetic disorder in Caucasians, with an incidence of approximately 1 in 3000 white live births (Kosorok et al., 1996). Chronic lung disease in children can be characterized by recurrent infections of *P. aeruginosa*, the most important pathogen that leads to the airway obstruction and lung function decay.

Pseudomonas aeruginosa infection was found to be a major predictor of morbidity and mortality (Kosorok et al., 2001). Young cystic fibrosis patients aged 1–5 years in 1990 with positive respiratory cultures for *P. aeruginosa* have significantly higher death rates and worse lung function during the following 8 years (Emerson et al., 2002). According to Li et al. (2005), about 30% of newborn infants acquired nonmucoid type of infection in the first 6 months of life, with a mucoid type of infection prevailing after age 4 years. It is of interest to characterize these patterns of infection in young cystic fibrosis patients.

The United States Cystic Fibrosis Foundation Patient Registry has documented the diagnosis and follow-up of all known cystic fibrosis patients from 114 accredited centres since the 1970s. The quality of this database improved greatly in 1986 because of more consistent reporting and quality control (FitzSimmons, 1993). In the 2007 registry data, there are 6585 subjects who were born after 1997 and have at least two follow-ups before the end of year 2007. The total length of follow-up is 27 412·7 person-years, averaging 4·2 years per subject. In these follow-up years, there were 10 353 nonmucoid and 3190 mucoid *P. aeruginosa* infections, along with 1339 events missing their category. Roughly, the occurrence rates are 3·8 for nonmucoid type and 1·2 for mucoid type per 10 years, not counting events with missing type. However, a patient may test positive for both nonmucoid and mucoid types at the same visit. To simplify the analysis, we treat the event with both types positive in the same visit as a third type of recurrent event process. Accordingly, there are 1582 such events during the follow-ups.

A large percentage of infections have missing category, so our estimation methods are preferable, as the complete case analysis that censors those events would have dramatic underestimation. Figure 2, derived by the rate proportion method and complete case analysis, reveals this. Particularly in nonmucoid type infections, there is substantial discrepancy between our estimates and the complete case analysis after the first year of age. In general, the two estimates diverge as age increases, partly due to more events with missing type being recorded over time. Based on the rate proportion method, the average number of nonmucoid type infections per patient is 2·4 by age 7, while that for mucoid type infections is 0·4. The rate for having both types of infections is similar to the rate for the mucoid type. Both increase more rapidly after age 7.

In Fig. 2, we also compare the estimation results between the rate proportion method and the weighted estimating equations method. We define the relative difference as the percent change of the weighted estimating equation estimates from the rate proportion estimates. In the weighted estimating equations approach, we used patient’s gender and mode of diagnosis as covariates.

The two methods produced similar results in estimating the nonmucoid *P. aeruginosa* infection rate with the relative difference being less than 5% over the range of the 10-year period. However, the infection rates of the mucoid type and of having both types in the same visit were significantly underestimated by the weighted estimating equation approach. The relative difference may reach as much as 50% in the first year of age.

6. Remarks

We assume that the observation probability $\pi_{ik}(t)$ is the same for each event type, which may not be realistic in practice when some types of events are more likely to have a missing category. However, the observation probability may not be estimable due to lack of information in those events with missing types. One possible generalization of our approach is to assume that the observation probability is known a priori for each category. One can show that, if $\pi_{ik}(t)$ is different for each k, our current approach leads to the estimation of

$p^*_{ik}(t)=E\{\delta_{ik}|\mathrm{d}N_{i\cdot}(t)=1, R_i(t)=1, Y_i(t)\}$, which differs from $p_{ik}(t) = E\{\delta_{ik} \mid \mathrm{d}N_{i\cdot}(t) = 1, R_i(t) = 0, Y_i(t)\}$, the unknown quantity in our mean function estimator. However, with $r_k(t)=\pi_{ik}(t)^{-1}p^*_{ik}(t)$ and $p_{ik}(t)=\{1-\pi_{ik}(t)\}r_k(t)/\sum_{\ell=1}^{K}\{1-\pi_{i\ell}(t)\}r_\ell(t)$, one may estimate $p_{ik}(t)$ with the estimation of $p^*_{ik}(t)$ and known $\pi_{ik}(t)$.

In the rate proportion method, the local likelihood procedure yields a nonparametric estimator via a regression model that uses time as a covariate. One may prefer to apply different non-parametric regression methods for categorized outcomes, such as the generalized additive model (Hastie & Tibshirani, 1990) or smoothing splines (Gu, 2002). It will be of interest to develop asymptotic theory for estimates based on such approaches and compare the performance across different nonparametric regression methods.

Acknowledgments

The authors thank Dr Preston Campbell from the United States Cystic Fibrosis Foundation for providing the registry data, and the editor and two referees for their helpful comments. This work is partially supported by the U.S. National Institutes of Health.

References

Abu-Libdeh H, Turnbull BW, Clark LC. Analysis of multi-type recurrent events in longitudinal studies: Application to a skin cancer prevention trial. Biometrics. 1990; 46:1017–34. [PubMed: 2085623]

Akaike H. A new look at the statistical model identification. IEEE Trans Auto Contr. 1974; 19:716–23.

Andersen, P.; Borgan, Ø.; Gill, R.; Keiding, N. Statistical Models Based on Counting Processes. New York: Springer; 1993.

Cai J, Schaubel D. Marginal means/rates models for multiple type recurrent event data. Lifetime Data Anal. 2004; 10:121–38. [PubMed: 15293628]

Chen BE, Cook RJ. The analysis of multivariate recurrent events with partially missing event types. Lifetime Data Anal. 2009; 15:41–58. [PubMed: 18622700]

Chiang CT, Wang MC, Huang CY. Kernel estimation of rate function for recurrent event data. Scand J Statist. 2005; 32:77–91.

Cook, RJ.; Lawless, JF. The Statistical Analysis of Recurrent Events. New York: Springer; 2007.

Cook RJ, Lawless JF, Lakhal-Chaieb L, Lee KA. Robust estimation of mean functions and treatment effects for recurrent events under event-dependent censoring and termination: Application to skeletal complications in cancer metastatic to bone. J Am Statist Assoc. 2009; 104:60–75.

Cook RJ, Lawless JF, Nadeau C. Robust tests for treatment comparisons based on recurrent event responses. Biometrics. 1996; 52:557–71. [PubMed: 8672703]

Emerson J, Rosenfeld M, McNamaral S, Ramsey B, Gibson RL. *Pseudomonas aeruginosa* and other predictors of mortality and morbidity in young children with cystic fibrosis. Pediatric Pulmonol. 2002; 34:91–100.

Fan, J.; Gijbels, I. Local Polynomial Modelling and Its Applications. London: Chapman & Hall; 1996.

FitzSimmons SC. The changing epidemiology of cystic fibrosis. J Pediatrics. 1993; 122:1–9.

Gu, C. Smoothing Spline ANOVA models. New York: Springer; 2002.

Hastie, T.; Tibshirani, R. Generalized Additive Models. London: Chapman & Hall; 1990.

Kosorok MR, Wei WH, Farrell PM. The incidence of cystic fibrosis. Statist Med. 1996; 15:449–62.

Kosorok MR, Zeng L, West SE, Rock MJ, Splaingard ML, Laxova A, Green CG, Collins J, Farrell PM. Acceleration of lung disease in children with cystic fibrosis after *Pseudomonas aeruginosa* acquisition. Pediatric Pulmonol. 2001; 32:277–87.

Lawless J, Nadeau C. Some simple robust methods for the analysis of recurrent events. Technometrics. 1995; 37:158–68.

Li Z, Kosorok MR, Farrell PM, Laxova A, West SEH, Green CG, Collins J, Rock MJ, Splaingard ML. Longitudinal development of mucoid *Pseudomonas aeruginosa* infection and lung disease progression in children with cystic fibrosis. J Am Med Assoc. 2005; 293:581–8.

Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. J R Statist Soc B. 2000; 62:711–30.

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. New York: Wiley; 2002.

Loader, C. R package version 1.5-6. 2010. locfit: Local Regression, Likelihood and Density Estimation.

Nelson WB. Graphical analysis of system repair data. J Qual Technol. 1988; 20:24–35.

Pepe MS, Cai J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. J Am Statist Assoc. 1993; 88:811–20.

Pollard, D. Empirical Processes: Theory and Applications. Hayward: Institute of Mathematical Statistics; 1990.

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. http://www.R-project.org

Schaubel D, Cai J. Rate/mean regression for multiple-sequence recurrent event data with missing event category. Scand J Statist. 2006; 33:191–207.

Appendix

We first provide the following regularity conditions.

Condition A1. Variables $\{N_{i1}(\cdot), \ldots, N_{iK}(\cdot)\}$ ($i = 1, \ldots, n$) are independent and identically distributed.

Condition A2. The expected number of subjects at risk $y_1(t)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\{Y_i(t)\}>0$ for every $t \in [0, \tau]$.

Condition A3. The total number of events $N_{i\cdot}(\tau) < \eta < \infty$.

Condition A4. For $t \in [0, \tau]$, observation probability $\pi_i(t) = E\{R_i(t)|dN_{ik}(t) = 1\}$ is the same for every k.

Condition A5. The likelihood function $\ell(\beta^*)$ is bounded and twice differentiable. The Hessian matrix $\ddot{\ell}(\beta^*) = \partial^2\ell(\beta^*)/\partial\beta^*\partial\beta^{*\mathrm{T}}$ is negative definite and invertible.

Condition A6. The function $\theta_k(\cdot)$ for each $k \in \{1, \ldots, K\}$ has a continuous $(q + 1)$th derivative for $q > 0$.

Condition A7. The kernel function $\kappa(\cdot)$ has a bounded and symmetric density with a compact support, and satisfies $\int \upsilon\, \kappa(\upsilon)\, d\upsilon = \int \upsilon^3\, \kappa(\upsilon)\, d\upsilon = 0$.

Condition A8. Assume $nh \to \infty$ as $h \to 0$ and $n \to \infty$.

Conditions A1–A3 are regularity conditions for recurrent event processes. We require data from the subjects to be independent and identically distributed in Condition A1. Our estimation, however, accommodates multiple dependent recurrent event processes. Condition A4 assumes that each type of event has the same probability for the category being missing. Conditions A5–A8 are otherwise regularity conditions for the large sample properties of the local likelihood estimates.

Proof of Theorem 1

To show the consistency of $\widehat{\mu}_k^r(t;\widehat{\theta})$, we decompose $\widehat{\mu}_k^r(t;\widehat{\theta})-\mu_k(t)$ as $\widehat{\mu}_k^r(t;\widehat{\theta})-\widehat{\mu}_k^r(t)+\widehat{\mu}_k^r(t)-\mu_k(t)$ with

$$\widehat{\mu}_k^r(t)=\widehat{\mu}_k^c(t)+\sum_{i=1}^{n}\int_0^t Y.(s)^{-1}\{1-R_i(s)\}p_k(s)\,\mathrm{d}N_{i\cdot}(s).$$

Let $\omega_k^{(1)}(t;\widehat{\theta})=\widehat{\mu}_k^r(t;\widehat{\theta})-\widehat{\mu}_k^r(t)$. First, we write

$$\omega_k^{(1)}(t;\widehat{\theta})=\sum_{i=1}^{n}\int_0^t Y.(s)^{-1}\{1-R_i(s)\}\{p_k(s;\widehat{\theta})-p_k(s)\}\,\mathrm{d}N_{i\cdot}(s).$$

Then, expanding $\omega_k^{(1)}(t;\widehat{\theta})$ around $\theta=(\theta_1, \ldots, \theta_{K-1})^{\mathrm{T}}$, we have

$$\omega_k^{(1)}(t;\widehat{\theta})=\sum_{i=1}^{n}\int_0^t Y.(s)^{-1}\{1-R_i(s)\}\Omega_k(s)^{\mathrm{T}}\{\widehat{\theta}(s)-\theta(s)\}\,\mathrm{d}N_{i\cdot}(s)+o(1).$$

It can be shown that $\omega_k^{(1)}(t;\widehat{\theta})$ converges in probability to

$$\int_0^t y_1(s)^{-1}\Omega_k(s)^{\mathrm{T}}e_v^{\mathrm{T}}b(s)\{1-\pi_1^*(s)\}r.(s)\,\mathrm{d}s.$$

Since the bias term $b(t)$ converges uniformly in probability to 0 when $h \to 0$, we can conclude that $\omega_k^{(1)}(t;\widehat{\theta})$ converges in probability to 0, uniformly in t. With $\widehat{\mu}_k^r(t)=\widehat{\mu}_k^n(t)+o(1)$ uniformly in t, we can prove the uniform consistency of $\widehat{\mu}_k^r(t;\widehat{\theta})$ by the fact that $\widehat{\mu}_k^n(t)$ uniformly converge to $\mu_k(t)$.

To prove the large sample normality we need to obtain the rate of the weak convergence when inserting in the local polynomial estimate. One can show that $\omega_k^{(1)}(t{:}\widehat{\theta})$ has the same weak convergence rate as

$$\int_0^t y_1(s)^{-1}\Omega_k(s)^{\mathrm{T}}\{\widehat{\theta}(s)-\theta_0(s)\}\{1-\pi_1^*(s)\}r.(s)\,\mathrm{d}s. \quad \text{(A1)}$$

Recall that the local polynomial estimate $\hat{\theta}(s)$ is $O(n^{-1/(2q+3)})$ when using the optimal bandwidth. Under the smoothness assumption of θ, one can show that (A1) is $O(n^{-(q+2)/(2q+3)})$, which is faster than $O(n^{-1/2})$. That means the sequence of $\omega_k^{(1)}(t;\widehat{\theta})$ will be dominated by the sequence of $\omega_k^{(2)}(t)=\widehat{\mu}_k^r(t)-\mu_k(t)$, which has a $O(n^{-1/2})$ weak convergence rate.

Combined with the asymptotic equivalency of $n^{1/2}\omega_k^{(2)}(t)$ and $n^{-1/2}\sum_{i=1}^{n}\int_0^t y_1(s)^{-1}\mathrm{d}M_{ik}^r(s)$, where $\mathrm{d}M_{ik}^r(t)=R_i(s)\,\mathrm{d}N_{ik}(s)+\{1-R_i(s)\}p_k(s)\,\mathrm{d}N_{i\cdot}(s)-Y_i(s)\,d\mu_k(s)$, one can show that $n^{1/2}\omega_k(t;\hat{\theta})$ is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}\varphi_{ik}(t;\theta_0)$, where

$$\varphi_{ik}(t;\theta)=\int_0^t y_1(s)^{-1}\Omega_k(s)^{\mathrm{T}}e_v^{\mathrm{T}}b(s)\{1-R_i(s)\}\,\mathrm{d}N_{i\cdot}(s)+\int_0^t y_1(s)^{-1}\mathrm{d}M_{ik}^r(s).$$

Notice that $\phi_{ik}(t;\theta_0)$ ($i = 1, \ldots, n$) are independent and identically distributed zero-mean variables, so $n^{-1/2}\sum_{i=1}^{n}\varphi_{ik}(t;\theta_0)$ converges to a multivariate normal distribution with mean zero and covariance $V_k(s, t) = E\{\phi_{1k}(s;\theta_0)\phi_{1k}(t;\theta_0)\}$ for $s, t \in [0, \tau]$. Hence $n^{1/2}\omega_k(t;\hat{\theta})$ converges weakly to a Gaussian process by the functional central limit theorem (Pollard, 1990), as the $\phi_{ik}(t;\theta_0)$ is composed of functions that are monotone in t, i.e., $\phi_{ik}(t;\theta_0)$ is manageable and $n^{1/2}\omega_k(t;\theta_0)$ is tight.

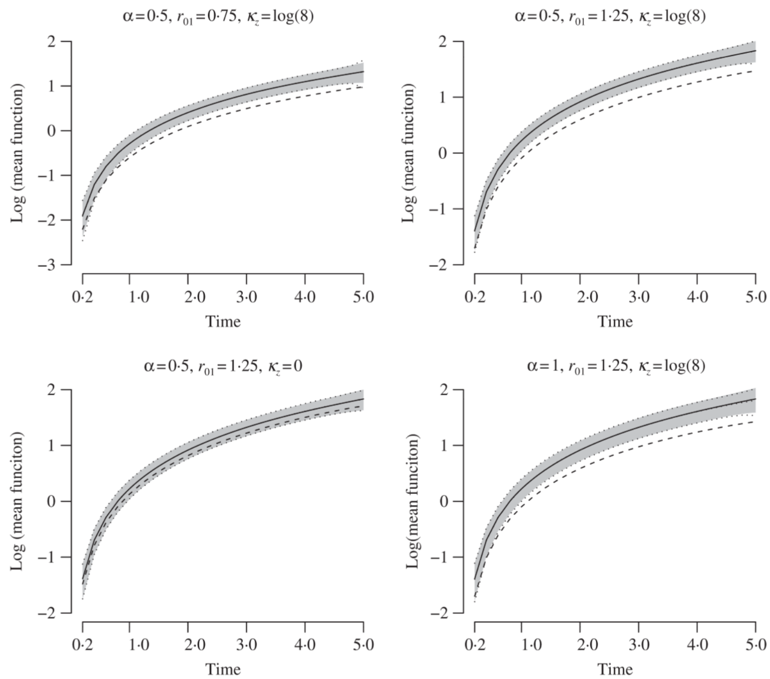

Fig. 1.

Mean function estimation of the truth (solid) by the rate proportion method with 95% confidence interval (dot) and complete case analysis (dash) under different simulation scenarios, with grey areas showing the truth ±1·96×(empirical standard errors).

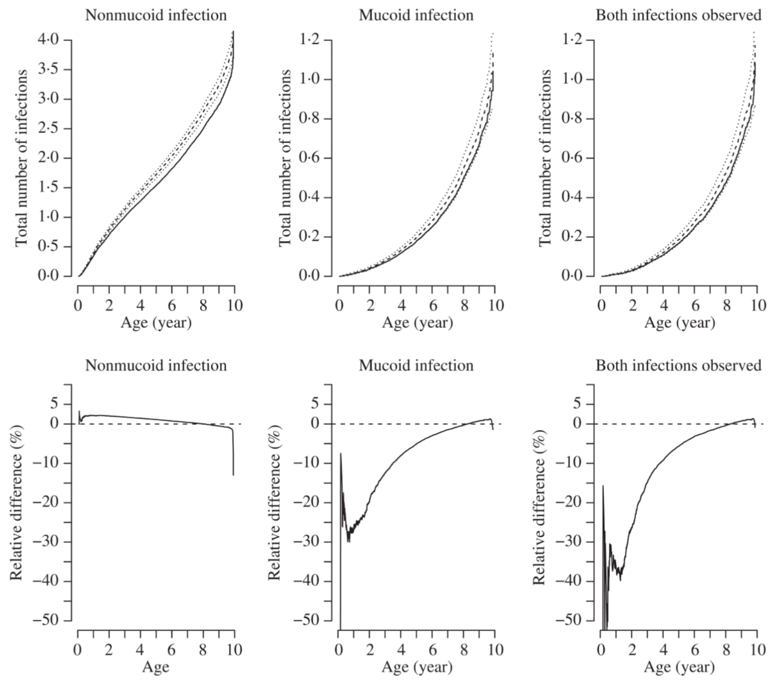

Fig. 2.
The upper panels show the mean function estimation of *P. aeruginosa* infections by the rate proportion method (dash) with 95% pointwise confidence interval (dot) and complete case analysis (solid); the lower panels present the relative difference between the rate proportion method and the weighted estimating equations method (solid) with the dashed line representing a reference line of zero difference.

Table 1

Simulation results for $\mu_1(t) = r_{01}t$ at $t = 3$; all entries except e_r^n and e_r^w are shown after multiplication by 10^2

α	$\mathbf{Bias}^n$	$\tilde V_1^{n1/2}$	κ_z	$\mathcal{M}_p$	$\mathbf{Bias}^r$	$\tilde V_1^{r1/2}$	$\tilde V_1^{r1/2}$	c_p	$\mathbf{Bias}^{\varpi}$	$\tilde V_1^{w1/2}$	e_r^n	e_r^w
$\mu_2(t) = r_{02}t^2/2$												
0·5	−0·45	17·9	0	10%	−0·13	18·0	18·2	94·5	−0·38	17·9	0·99	0·99
				20%	0·09	18·3	18·1	94·6	−0·38	18·1	0·96	0·98
				30%	0·25	18·7	18·0	94·0	−0·48	18·4	0·91	0·97
			2·08	10%	−0·18	18·3	18·1	93·8	−0·45	18·2	0·96	0·99
				20%	0·00	18·6	18·1	93·7	−0·46	18·3	0·93	0·98
				30%	0·49	18·9	18·1	92·9	−0·34	18·6	0·90	0·97
1·0	0·93	22·8	0	10%	1·05	22·9	22·2	93·8	0·86	22·9	0·99	0·99
				20%	1·30	23·3	22·2	93·8	0·85	23·1	0·96	0·99
				30%	1·79	23·4	22·3	94·2	1·10	23·2	0·94	0·98
			2·08	10%	1·12	23·1	22·3	93·6	0·95	23·0	0·98	0·99
				20%	1·42	23·1	22·3	94·2	0·99	23·0	0·97	0·99
				30%	1·84	23·6	22·3	94·0	1·10	23·3	0·93	0·97
$\mu_2(t) = r_{02}(t + t^2/2 + t^3/9)$												
0·5	0·52	18·3	0	10%	0·17	18·6	18·1	93·8	2·04	18·7	0·97	1·03
				20%	0·03	19·0	17·9	92·6	3·21	19·2	0·92	1·05
				30%	−0·06	19·7	17·8	91·5	4·56	19·8	0·86	1·07
			2·08	10%	0·29	18·5	18·1	93·7	1·81	18·7	0·98	1·03
				20%	0·11	19·0	17·9	92·9	2·92	19·3	0·93	1·05
				30%	0·52	19·7	17·8	91·9	4·72	20·4	0·86	1·13
1·0	2·07	21·8	0	10%	1·63	22·0	22·1	94·6	4·90	22·5	0·98	1·09
				20%	1·35	22·3	21·9	94·4	6·62	23·1	0·96	1·16
				30%	1·20	22·8	21·8	93·7	8·55	23·8	0·91	1·22
			2·08	10%	1·59	21·9	22·2	94·9	4·32	22·4	0·99	1·08
				20%	1·32	22·3	22·0	94·4	6·21	23·1	0·95	1·14
				30%	1·35	22·7	21·9	94·1	8·13	23·8	0·92	1·23

Table 2

Simulation results for $\mu_1(t) = r_{01}t$ when $K = 3$ and the frailty follows a log-normal distribution; all entries except e_r^n and e_r^w are shown after multiplication by 10^2

t	n	Bias^n	$\tilde{V}^{n1/2}$	κ_z	$\mathcal{M}_p$	Bias^r	$\tilde{V}^{r1/2}$	$V^{r\bar{1}/2}$	$\mathcal{C}_p$	Bias$^\varpi$	$\tilde{V}^{\varpi 1/2}$	e_r^n	e_r^w
1	50	0·83	14·5	0	15%	0·76	14·7	16·1	92·4	0·46	14·6	0·99	0·98
					25%	0·63	14·9	18·1	93·8	0·09	14·7	0·96	0·98
					35%	0·65	15·2	20·7	94·3	−0·16	14·9	0·92	0·96
				2·08	15%	0·80	14·7	16·3	92·0	0·47	14·6	0·97	0·98
					25%	0·76	15·0	18·1	93·9	0·19	14·8	0·95	0·97
					35%	0·68	15·2	20·9	94·5	−0·20	14·9	0·92	0·96
	200	0·24	7·11	0	15%	0·35	7·24	7·78	95·7	−0·13	7·19	0·96	0·98
					25%	0·36	7·31	8·39	96·3	−0·51	7·20	0·94	0·97
					35%	0·52	7·39	9·14	96·8	−0·81	7·24	0·92	0·97
				2·08	15%	0·32	7·20	7·84	95·3	−0·21	7·15	0·97	0·98
					25%	0·40	7·27	8·46	96·4	−0·51	7·18	0·95	0·98
					35%	0·48	7·46	9·36	97·1	−0·95	7·30	0·90	0·97
3	50	3·66	38·6	0	15%	4·45	38·2	38·9	92·1	5·47	39·4	1·02	1·07
					25%	5·10	38·3	40·3	92·9	6·33	40·0	1·01	1·10
					35%	5·54	38·3	42·1	92·5	6·86	40·2	1·01	1·11
				2·08	15%	4·49	38·4	39·0	91·6	5·16	39·4	1·01	1·06
					25%	5·03	38·4	40·4	92·0	5·92	39·9	1·00	1·08
					35%	6·58	39·5	42·7	93·1	7·55	41·6	0·94	1·11
	200	0·48	19·0	0	15%	1·17	19·0	18·8	93·1	2·28	19·6	1·00	1·07
					25%	1·76	19·3	19·2	94·0	3·13	20·0	0·96	1·09
					35%	2·33	19·3	19·7	94·0	3·81	20·1	0·96	1·10
				2·08	15%	1·28	19·3	18·9	93·9	2·17	19·8	0·97	1·06
					25%	1·83	19·3	19·3	94·5	2·88	19·9	0·97	1·08
					35%	2·88	19·3	20·0	95·1	4·00	20·3	0·95	1·12