ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)
66, *Part* 4, *pp.* 833–846

A second-order semiparametric method for survival analysis, with application to an acquired immune deficiency syndrome clinical trial study

Fei Jiang,

University of Hong Kong, People's Republic of China

Yanyuan Ma

Penn State University, State College, USA

and J. Jack Lee

University of Texas MD Anderson Cancer Center, Houston, USA

[Received August 2015. Final revision August 2016]

Summary. Motivated by the recent acquired immune deficiency syndrome clinical trial study A5175, we propose a semiparametric framework to describe time-to-event data, where only the dependence of the mean and variance of the time on the covariates are specified through a restricted moment model. We use a second-order semiparametric efficient score combined with a non-parametric imputation device for estimation. Compared with an imputed weighted least squares method, the approach proposed improves the efficiency of the parameter estimation whenever the third moment of the error distribution is non-zero. We compare the method with a parametric survival regression method in the A5175 study data analysis. In the data analysis, the method proposed shows a better fit to the data with smaller mean-squared residuals. In summary, this work provides a semiparametric framework in modelling and estimation of survival data. The framework has wide applications in data analysis.

Keywords: CD4 cell counts; Censoring; Efficiency; Imputation; Kernel; Non-parametric methods; Restricted moments; Safety end points; Two-stage analysis; Toxicity

1. Introduction

A new Acquired Immune Deficiency Syndrome Clinical Trials Group study, A5175, was recently conducted to evaluate several antiretroviral regimens in diverse populations. One primary goal of the study is to investigate the safety of these regimens to maximize the efficiency of the antiretroviral delivery in various areas (Campbell *et al.*, 2012). The primary safety end point of the study is a patient's time to one of the following three early adverse reactions: onset of a grade 3 or higher severity sign, a grade 3 or higher laboratory abnormality and a change of the initial treatment due to toxicity of the treatment. A patient's event was considered to be censored if he or she did not meet the primary end point criteria at the end of the study or at the final medication dose. In addition, the study also collected patients' CD4 cell counts at the baseline and then at weeks 8, 24, 72 and 96. Compared with the primary safety end point, the CD4 cell counts information was obtained relatively easily in a shorter period of time. Although the CD4

Address for correspondence: Fei Jiang, University of Hong Kong, Hong Kong, People's Republic of China.
E-mail: feijiang@hku.hk

 0035–9254/17/66833

cell counts information is primarily used in inferring the treatment efficacy (Campbell *et al.*, 2012), it is also related to the safety of the antiretroviral regimens. For example, Hirsch (2008) showed that using the same antiretroviral regimen at a higher CD4 cell count level would lower the risk of toxicities. Thus, it is natural to expect that an analysis on the primary safety end point would be more efficient if the short-term information on CD4 cell counts can be included. This motivates us to develop methods to analyse the relationship between CD4 cell counts and the primary safety end point, with the goal of ameliorating the existing post-trial data analysis procedures. In addition, we also explore the usage of the proposed methods in the clinical trial design stages to improve trial efficiency.

In the A5175 study, safety of a treatment is described by time to adverse events, and all the subsequent decisions are made on the basis of the inference on the event time. This motivates us to model the time to the primary safety end point directly as a function of the covariates. In contrast, traditional time-to-event models such as the Cox proportional hazard model focus on evaluating the covariate effect on the risk of disease and do not provide direct inference on the event time. Our preliminary analysis (Section 3.3) on the A5175 study data shows that both the mean and the variance of the primary safety end point depend on the short-term CD4 cell counts. To capture this relationship while remaining flexible, we use a semiparametric second-order restricted moment model to specify the mean and variance structures of the primary safety end point while leaving all other aspects of the model unspecified. The model has the characteristics of capturing the central structure while remaining flexible in non-crucial parts of the model. By modelling the variance in addition to the mean, the second-order restricted moment model enriches the structure of the classical restricted moment model.

To obtain accurate parameter estimation and to perform proper inference on time to the primary safety end point, we devise a semiparametric estimation procedure for the second-order restricted moment model used in fitting the A5175 data. To our best knowledge, such modelling and estimation approaches have not been considered in survival models. In classical regression models, parameter estimation is often performed by using the ordinary least squares method, which is efficient when the errors are normally distributed (Gallant, 2009). However, the additional variance structures in the A5175 study data imply that the ordinary least squares estimators may not be optimal. Under the complete-data settings, Wang and Leblanc (2008) proposed a second-order least squares method when the error variances are constant. The method was later generalized to covariate-dependent error variances and shown to minimize the variances of the estimators (Kim and Ma, 2012).

The A5175 study data are further subject to censoring. This prevents the direct application of the methods that were described above because, without fully specifying the event time distribution, the score functions of the censored subjects are difficult to obtain. In a completely different context, Wang *et al.* (2012) proposed a non-parametric score imputation method to cope with censoring when covariates are discrete. The non-parametric score imputation method often performs competitively compared with the optimal augmented inverse probability weighting method in terms of estimation variability in finite samples (Wang *et al.*, 2012), whereas the former has a more intuitive form and is more interpretable. This inspires us to examine the non-parametric imputation strategy and to extend the method to incorporate continuous covariates (CD4 cell counts in the A5175 study data). We then generalize the semiparametric estimation method of Kim and Ma (2012) to handle survival data. We develop an imputation-based semiparametric efficient estimator for the second-order restricted moment model, RMM2-ISE, which combines the non-parametric score imputation with the second-order least squares score function that was introduced in Kim and Ma (2012). We derive its asymptotic estimation variance and establish its root n consistency and asymptotic normality. We evaluate the finite sample

properties of RMM2-ISE. We further compare the RMM2-ISE estimation procedure with a simpler method, which we name the imputed weighted least squares (IWLS) method, through simulation studies. We developed IWLS here to combine non-parametric score imputation and weighted least squares score functions. A similar idea was used in Lipsitz *et al.* (1999) to handle missing covariates. Moreover, we apply the RMM2-ISE-method to analyse the A5175 study data. The RMM2-ISE-method also shows better data fitting compared with the method combining the accelerated failure time (AFT) Weibull model and maximized likelihood (ML) estimation. Throughout the paper, we choose the Weibull survival time model to fit the data for comparisons because it is sufficiently flexible to accommodate increasing, decreasing and constant hazard rates (Klein and Moeschberger, 2010).

The rest of the paper is structured as follows. In Section 2, we describe the second-order restricted moment model and introduce a second-order semiparametric efficient estimator. We also describe the non-parametric imputation method for treating censored observations and study its properties. In Section 3, we analyse the A5175 study data by using our modelling and estimation methods, after examining them via simulation studies. We conclude the paper with a discussion in Section 4 and relegate all the technical proofs to the on-line supplementary document.

2. Modelling and methodological development

2.1. Second-order restricted moment model in complete data

We first introduce the second-order restricted moment model under the general complete-data settings; we then define the specific model for the A5175 study data. Let Y_i and W_i denote the independently and identically distributed response random variables and covariates respectively. In our paper, Y_i is the survival time on the logarithmic scale. Let $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote the parameters that are associated with the mean and the variance respectively. A general second-order restricted moment model has the form

$$g(Y_i) = m(W_i, \boldsymbol{\beta}) + \xi_i, \tag{1}$$

where $g(\cdot)$ is a known link function, $E(\xi_i|W_i) = 0$ and $E(\xi_i^2|W_i) = \sigma^2(W_i, \boldsymbol{\gamma})$. Here $m(\cdot)$ is a generic function known up to the parameter $\boldsymbol{\beta}$ and $\sigma^2(\cdot)$ is a generic positive function known up to the parameter $\boldsymbol{\gamma}$. Differently from the usual regression models, the error variation is also specified as a function of W_i.

Based on Kim and Ma (2012), the semiparametric efficient estimator can be obtained by solving estimating equations that are formed by the sum of the efficient score functions

$$\begin{aligned} \mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}(W_i, Y_i) &= \frac{\partial m(W_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left\{ \frac{\xi_i}{\sigma^2(W_i, \boldsymbol{\gamma})} - \frac{E(\xi_i^3|W_i) D_i}{\sigma^2(W_i, \boldsymbol{\gamma}) E(D_i^2|W_i)} \right\} \\ \mathbf{S}_{\boldsymbol{\gamma},\mathrm{eff}}(W_i, Y_i) &= \frac{D_i}{E(D_i^2|W_i)} \frac{\partial \sigma^2(W_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}, \end{aligned} \tag{2}$$

where

$$D_i = \xi_i^2 - \sigma^2(W_i, \boldsymbol{\gamma}) - E(\xi_i^3|W_i)\xi_i/\sigma^2(W_i, \boldsymbol{\gamma}).$$

When the third moment $E(\xi_i^3|W_i) = 0$, the score function for $\boldsymbol{\beta}$ is the same as that for the ordinary least squares estimator. This fact shows that, in estimating $\boldsymbol{\beta}$, the resulting estimator is at least as efficient, and is often more efficient, compared with the ordinary least squares estimator. Further, if $E(\xi_i^3|W_i) \neq 0$, the resulting estimator gains efficiency by making use of the

additional variance structure. We point out that, although the true third and fourth moments of ξ_i conditional on W_i are needed in expression (2), in practice, their parametrically or non-parametrically estimated versions can be plugged in and the resulting estimation efficiency of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ will not be affected (Kim and Ma, 2012). In Section 3, we provide specific estimators of $E(\xi_i^3|W_i)$ and $E(\xi_i^4|W_i)$ both parametrically and non-parametrically, using no additional data.

The above efficient score functions are derived under the complete-data setting. In the next section, we modify the efficient score functions and introduce the estimating equations for censored survival data. We further derive the statistical properties of the resulting estimators.

2.2. The imputation estimator

The A5175 study data are complicated by censoring. More specifically, let T_i and C_i be the primary safety end point and the censoring time for the ith subject on the logarithmic scale. We observe only $X_i = \min(T_i, C_i)$ and the censoring indicator $\Delta_i = I(T_i \leqslant C_i)$, for $i = 1, \ldots, n$. A widely accepted method for handling censoring is the likelihood-based approach, such as that used for the AFT–Weibull model. Because of the full parameterization of the survival time distribution in AFT–Weibull models, the probability that an event happens after a certain time can be expressed as a function of a finite dimensional parameter. The parameter estimation can then be performed through maximizing the likelihood of the observed data. Although this method has long been known, its application is limited because of its non-robustness, in that, as soon as the true population distribution deviates from the AFT–Weibull model, the method leads to misleading results. In this paper, we introduce a non-parametric score imputation method to deal with the censored primary safety end points, which makes much fewer assumptions and is more robust. The method extends the approach of Wang *et al.* (2012) under the discrete setting by including the CD4 cell counts as a continuous covariate. Combined with the second-order restricted moment model and the semiparametric efficient score equations, the method yields consistent estimators as long as the first two moment assumptions are satisfied.

Throughout the text, we use capital letters to denote a random variable and small letters to denote the corresponding realizations. For identifiability and simplicity, we assume that the censoring distribution is independent of the survival time and the covariates. We consider the efficient score function $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, t_i) = (\mathbf{S}_{\boldsymbol{\beta},\mathrm{eff}}(w_i, t_i)^{\mathrm{T}}, \mathbf{S}_{\boldsymbol{\gamma},\mathrm{eff}}(w_i, t_i)^{\mathrm{T}})^{\mathrm{T}}$ for the parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, where w_i and t_i are the values of the CD4 counts and the primary safety end point respectively. We define the RMM2-ISE estimating equation under the survival settings as

$$\sum_{i=1}^{n} \delta_i \mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, t_i) + (1 - \delta_i) E\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, T_i) | T_i > X_i, W_i = w_i, X_i = x_i\}, \tag{3}$$

where δ_i is the realization of Δ_i. Thus, if a subject has an observed primary safety end point, we use the original efficient score function. However, if a subject is censored, we use the expected value of the score function conditional on the CD4 cell counts, given that no adverse reaction has happened before the censoring time.

Without specifying the population distribution of the primary safety end point, we evaluate the conditional expectation in model (3) non-parametrically via kernel methods, which has good asymptotic properties with a properly chosen bandwidth (Devroye, 1981). We define

$$\begin{aligned} \mathbf{Q}_{\boldsymbol{\theta},i}(w_i, x_i) &= E\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, T_i) | T_i > X_i, W_i = w_i, X_i = x_i\} \\ &= E\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, T_i) | T_i > x_i, W_i = w_i, C_i = x_i\} \end{aligned}$$

$$= \frac{E\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, T_i) I(T_i > x_i) | W_i = w_i, C_i = x_i\}}{E\{I(T_i > x_i) | W_i = w_i, C_i = x_i\}}$$
$$= \frac{E\{\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, T_i) I(T_i > x_i) | W_i = w_i\}}{E\{I(T_i > x_i) | W_i = w_i\}},$$

where the last equality is because C_i and T_i are independent given W_i. If T_is are observed, we would simply use non-parametric kernel regressions to approximate the two conditional expectations above. However, because T_is are observed only when $\Delta_i = 1$, we further need to modify the two averages with the inverse-probability-weighted averages, where the weights are the probability of censoring time after event time, i.e. the survival function of the censoring process $G(\cdot|W) = G(\cdot)$ under the assumption that censoring is independent of the covariate. The kernel estimator of $\hat{\mathbf{Q}}_{\boldsymbol{\theta},i}$ is thus written as

$$\hat{\mathbf{Q}}_{\boldsymbol{\theta},i}(w_i, x_i) = \frac{\sum_{j=1}^{n} \delta_j \mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_j, x_j) I(x_j > x_i) K_h(w_j - w_i) / \hat{G}(x_j)}{\sum_{j=1}^{n} \delta_j I(x_j > x_i) K_h(w_j - w_i) / \hat{G}(x_j)}, \tag{4}$$

where

$$\hat{G}(t_j) \equiv \prod_{x_i \leqslant t_j} \left\{ 1 - \frac{(1 - \Delta_i)}{\sum_{k=1}^{n} I(x_k \geqslant x_i)} \right\}$$

is the Kaplan–Meier estimator for the survival function of the censoring distribution $G(\cdot)$, and $K_h(\cdot) \equiv K(\cdot/h)/h$, where K is a kernel function and h is a bandwidth. When $h \to 0$, the imputed score functions reduce to those introduced in Wang *et al.* (2012) in the discrete covariate settings.

Specifically, to obtain $\hat{\mathbf{Q}}_{\boldsymbol{\theta},i}(w_i, x_i)$, we use the product limit estimator to estimate G. We choose the Gaussian kernel with bandwidth $h = n^{-2/15} h_{\mathrm{s}}$, where $h_{\mathrm{s}} = 1.06\sigma n^{-1/5}$ is Silverman's rule-of-thumb bandwidth (Silverman (1986), page 45) and σ is the standard deviation of W_i. Because h_{s} has order $n^{-1/5}$, the proposed bandwidth h satisfies $nh^4 \to 0$ and $nh^2 \to \infty$ when $n \to \infty$. Because of the indicators δ_j and $I(x_j > x_i)$, only the uncensored data from the individuals who have not met the safety event criteria at x_i contribute to the summations in $\hat{\mathbf{Q}}_{\boldsymbol{\theta},i}(w_i, x_i)$. After computing $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_j, t_j)$ and $\hat{\mathbf{Q}}_{\boldsymbol{\theta},i}(w_i, x_i)$ for the uncensored and censored observations respectively, we obtain RMM2-ISE $\hat{\boldsymbol{\theta}}$ through solving the estimating equation

$$\sum_{i=1}^{n} \delta_i \mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}(w_i, t_i) + (1 - \delta_i) \hat{\mathbf{Q}}_{\boldsymbol{\theta},i}(w_i, x_i) = 0. \tag{5}$$

Under assumptions A1–A8 listed in on-line appendix A.1, we rigorously establish the consistency and asymptotic properties of the estimator, i.e. we obtain $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = o_p(1)$, and $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N\{0, A^{-1}\Omega(A^{-1})^{\mathrm{T}}\}$ in distribution, where A and Ω are defined in theorem 2 of the on-line appendix. We elaborate the consistency and asymptotic normality of RMM2-ISE in theorems 1 and 2 followed by their detailed proofs in the appendix in the on-line supplementary document.

3. Analysis of the A5175 study data

We are now ready to analyse the A5175 study data by using the RMM2-ISE-method. Before the analysis, we first perform a numerical evaluation of the estimation procedure on simulated samples and compare the estimation results with the IWLS method that was introduced in

Section 1. The IWLS estimator is obtained by solving equation (5), but with $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$ in it replaced by $\sigma^{-2}(W_i)\xi_i\partial m(W_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, which is the score function that is associated with the weighted least squares method. Here $\sigma^2(W_i)$ is the conditional variance of ξ_i given W_i, which can be replaced by its consistent estimator. The consistent estimator can be obtained by using the non-censored observations, because our score function is first constructed for the fully observed samples which relies on the $\sigma^2(W_i)$ for only the non-censored observations. We discuss several ways of estimating $\sigma^2(W_i)$ later in this section. Note that the same replacement of $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$ is needed in calculating $\hat{\mathbf{Q}}_{\boldsymbol{\theta},i}(w_i,x_i)$ in equation (5). The asymptotic variance of the IWLS estimator can be shown to be the same as Ω in theorem 2, except that $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$ needs to be replaced by $\sigma^{-2}(W_i)\xi_i\partial m(W_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\mathbf{Q}_{\boldsymbol{\theta}_0,i}$ is also adapted correspondingly. It is readily seen that the asymptotic estimation variances of the RMM2-ISE- and the IWLS methods have the same structure except for the different forms of $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$. This suggests that, intuitively, the RMM2-ISE-method would have better asymptotic efficiency, because the score function for RMM2-ISE is more efficient than that for IWLS (Wang and Leblanc, 2008; Kim and Ma, 2012) in the complete-data settings, and the kernel imputation induces the same type of asymptotic variance inflation for both methods when the data are subject to censoring. We explore the required sample sizes and censoring rates for implementing the RMM2-ISE-procedure and show that the procedure yields accurate estimators under reasonable uncensored sample sizes. Moreover, we show via simulation that the RMM2-ISE-method gains efficiency compared with the simpler IWLS method when the third moment of the error distribution does not vanish. These conclusions are crucial, because they support the applications of the RMM2-ISE-method to the A5175 study data.

3.1. Evaluation of methods

We illustrate the relative performance of RMM2-ISE and the IWLS estimator through demonstrating that the former is more efficient than the latter. Note that we use the same imputation method in both estimation procedures.

In the complete-data setting, RMM2-ISE is shown to be more efficient than the IWLS estimator when the conditional third moment of the error distribution is non-zero (Wang and Leblanc, 2008; Kim and Ma, 2012). To illustrate this point as well as the consistency of the estimators under the setting with censoring, we generate the data as follows. The covariate W_i is the logarithm of a random variable generated from the uniform (0, 5) distribution. The error term $\xi_i = \chi^2(k_i) - k_i$, where $\chi^2(k_i)$ is generated from the χ^2-distribution with degrees of freedom $k_i = (\gamma_0 + \gamma_1 W_i^2)/2$. Note that the variance of ξ_i is $\sigma^2(W_i,\boldsymbol{\gamma}) = 2k_i$, which depends on the covariate, and $E(\xi_i^3|W_i)$ does not vanish. We generate the time-to-safety end point T_i from the exponential model

$$\log(T_i) = \beta_0 \exp(\beta_1 W_i) + \xi_i. \tag{6}$$

We further generate the censoring time from exponential distributions. We vary the exponential rate parameters to obtain various censoring rates. We assess the performances of RMM2-ISE and the IWLS estimator at various censoring rates.

Following model (2), we obtain the semiparametric efficient score functions for the above model as

$$\mathbf{S}_{\boldsymbol{\beta},\text{eff}}(W,T) = (\exp(\beta_1 W), \beta_0 \exp(\beta_1 W) W)^{\mathrm{T}} \left\{ \frac{\xi}{\sigma^2(W,\boldsymbol{\gamma})} - \frac{E(\xi^3|W)D}{\sigma^2(W,\boldsymbol{\gamma})E(D^2|W)} \right\},$$

$$\mathbf{S}_{\boldsymbol{\gamma},\text{eff}}(W,T) = (1, W^2)^{\mathrm{T}} \frac{D}{E(D^2|W)}.$$

Table 1. Comparisons of the optimal IWLS estimator and RMM2-ISE†

Truth		*Results for IWLS*				*Results for RMM2-ISE*					
β_0	β_1	$\hat{\beta}_0$	$\hat{\beta}_1$	$SD(\hat{\beta}_0)$	$SD(\hat{\beta}_1)$	$\hat{\beta}_0$	$\hat{\beta}_1$	$SD(\hat{\beta}_0)$	$SD(\hat{\beta}_1)$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
0% censoring rate											
1.0	−0.2	0.954	−0.195	0.053	0.041	0.972	−0.196	0.048	0.036	0.802	0.071
1.0	−0.4	0.958	−0.393	0.056	0.033	0.974	−0.399	0.046	0.030	0.794	0.078
1.0	−0.6	0.966	−0.593	0.060	0.034	0.980	−0.598	0.049	0.030	0.797	0.091
1.0	−0.8	0.970	−0.790	0.070	0.040	0.988	−0.793	0.054	0.037	0.840	0.097
15% censoring rate											
1.0	−0.2	0.895	−0.187	0.053	0.045	0.951	−0.190	0.042	0.041	0.728	0.068
1.0	−0.4	0.894	−0.391	0.057	0.043	0.951	−0.392	0.045	0.037	0.730	0.078
1.0	−0.6	0.897	−0.595	0.067	0.047	0.956	−0.588	0.048	0.041	0.742	0.084
1.0	−0.8	0.894	−0.803	0.074	0.061	0.957	−0.786	0.055	0.048	0.751	0.093
25% censoring rate											
1.0	−0.2	0.851	−0.184	0.057	0.049	0.925	−0.180	0.046	0.049	0.744	0.063
1.0	−0.4	0.850	−0.396	0.059	0.047	0.926	−0.385	0.047	0.044	0.742	0.067
1.0	−0.6	0.849	−0.607	0.066	0.057	0.929	−0.587	0.055	0.053	0.755	0.068
1.0	−0.8	0.842	−0.823	0.073	0.073	0.927	−0.787	0.059	0.062	0.768	0.078
50% censoring rate											
1.0	−0.2	0.736	−0.193	0.057	0.053	0.848	−0.185	0.054	0.054	0.794	0.042
1.0	−0.4	0.729	−0.425	0.065	0.057	0.855	−0.386	0.055	0.059	0.776	0.059
1.0	−0.6	0.722	−0.665	0.075	0.082	0.853	−0.605	0.063	0.073	0.787	0.067
1.0	−0.8	0.709	−0.912	0.088	0.122	0.848	−0.814	0.071	0.097	0.792	0.070
75% censoring rate											
1.0	−0.2	0.581	−0.248	0.055	0.074	0.739	−0.218	0.061	0.073	0.843	0.034
1.0	−0.4	0.577	−0.558	0.063	0.123	0.741	−0.465	0.058	0.092	0.838	0.033
1.0	−0.6	0.567	−0.889	0.072	0.200	0.736	−0.711	0.072	0.138	0.818	0.048
1.0	−0.8	0.556	−1.245	0.080	0.309	0.724	−0.953	0.076	0.191	0.795	0.069

†Sample size, $n = 400$; $\beta_0 = 1$; $\gamma = (1, 0.1)^{\mathrm{T}}$. SD represents the sample empirical standard deviation based on 1000 simulations.

We then impute the above score functions as described in Section 2 to estimate the parameters.

We use the true $E(\xi_i^3|W_i)$ and $E(\xi_i^4|W_i)$ to obtain RMM2-ISE and use the true $E(\xi_i^2|W_i)$ to form the optimal weights $1/\sigma^2(W_i, \gamma_0)$ to obtain the IWLS estimator. This guarantees that both estimators achieve their optimal performance in the complete-data setting. In other words, we avoid the hidden efficiency loss due to possible misspecification of moment functions in both estimators to keep the comparison fair. We compare the biases and variances of the resulting RMM2-ISE and IWLS estimators in all the numerical experiments.

3.2. Numerical results for the estimation procedures

We use a sample size $n = 400$ and generate 1000 data sets from model (6), with $\beta_0 = 1$ and $\gamma = (1, 0.1)$. In Table 1, we present the performance of RMM2-ISE and the IWLS estimator under different specifications of β_1 and censoring rates. Here $E(\xi_i^3|W_i)$ is estimated through fitting a linear model between $\hat{\xi}_i^3$ and the covariates, and $E(\xi_i^4|W_i)$ is estimated through fitting a quadratic model between $\hat{\xi}_i^4$ and the covariates. Here $\hat{\xi}_i$s are the residuals after fitting a linear regression for the non-censored observations. The linear model is simple and the most common

Table 2. Estimation variations when $\beta_0 = 1, \beta_1 = -0.6, \gamma_0 = 1$ and $\gamma_1 = 0.1$†

Censoring rate (%)	$SD(\hat{\beta}_0)$	$SD(\hat{\beta}_1)$	$\widehat{SD}(\hat{\beta}_0)$	$\widehat{SD}(\hat{\beta}_1)$	$SD(\hat{\gamma}_0)$	$SD(\hat{\gamma}_1)$	$\widehat{SD}(\hat{\gamma}_0)$	$\widehat{SD}(\hat{\gamma}_1)$
0	0.049	0.030	0.045	0.031	0.171	0.095	0.254	0.118
15	0.048	0.041	0.054	0.039	0.128	0.100	0.174	0.124
25	0.055	0.053	0.045	0.036	0.127	0.088	0.161	0.123
50	0.063	0.073	0.028	0.028	0.137	0.071	0.102	0.075

†SD represents the empirical standard deviation from the 1000 simulation runs. $\widehat{\text{SD}}$ represents the theoretic asymptotic standard deviation.

regression model in practice, whereas the quadratic model ensures the non-negativeness of the regression function. We first fit the working model on the basis of the non-censored residuals and covariates, and then use the fitted model to impute the additional censored moments. Neither the linear nor the quadratic model is the true model of these conditional moments. However, for the IWLS method, we used $E(\xi_i^2|W_i)$ under the true model. This means that we compared a sub-optimal RMM2-ISE-method with the optimal IWLS method. Hence theoretically there is no guarantee that RMM2-ISE should outperform the IWLS estimator. We used this particularly harsh setting for RMM2-ISE to test its performance stability and robustness to the working models. As we can see, if no observation is censored (the censoring rate is 0%), both estimates are close to the true values. This illustrates the consistency of the estimators when no observation is censored. Further, RMM2-ISE has smaller biases and variances compared with the IWLS estimator, which illustrates the better accuracy and efficiency of RMM2-ISE compared with the IWLS estimator. When the censoring rate is greater than 0, RMM2-ISE continues to perform well. In fact, even when the censoring rate is moderately large (25%), RMM2-ISE is still close to the truth (with less than 0.1 absolute biases). Because censoring reduces the information that is contained in the sample for inferring the population distributions, both RMM2-ISE and the IWLS estimator start to deteriorate when the censoring rates further increase. However, RMM2-ISE has smaller deterioration compared with the IWLS estimator under all situations. For example, the IWLS estimator for β_0 shows more than 0.1 absolute biases when the censoring rate is 15%, whereas the corresponding RMM2-ISE-estimator keeps the absolute biases within 0.1 until the censoring rate reaches 50%. Compared with the estimation of β_0, the RMM2-ISE- and IWLS methods perform better in estimating the parameter of clinical interest, β_1. Nevertheless, the IWLS estimator has biases greater than 0.1 when the censoring rate is 50%, whereas this occurs for RMM2-ISE only when the censoring rate reaches 75%. Overall, compared with the IWLS estimator, RMM2-ISE generally has smaller biases in estimating $\boldsymbol{\beta}$. The standard deviations of RMM2-ISE are smaller than those of the IWLS estimator on average. In conclusion, the RMM2-ISE-method performs better than the IWLS method in terms of smaller biases and variations in the resulting estimation. In the simulation studies, we see that the bias increases when the censoring rate increases. Compared with $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ has larger bias and variance. However, this does not indicate that the estimator is inconsistent. In fact, when we further increase the sample size, we observe a clear reduction in the biases. Thus, the relatively large bias at high censoring rate that we observe here is a finite sample phenomenon.

In Table 2, we compare the estimated asymptotic standard deviation derived in theorem 2 with the empirical estimation standard deviation summarized from the simulated samples. The results show that, when the censoring rate is small (25% or less), the asymptotic standard deviation estimators are close to the empirical estimators, whereas their performance deteriorates when

Table 3. Estimation results from 1000 simulation runs when $n = 800$, $\beta_0 = 1$, $\beta_1 = -0.6$, $\gamma_0 = 1$ and $\gamma_1 = 0.1$†

Censoring rate (%)	*Results for IWLS*				*Results for RMM2-ISE*					
	$\hat{\beta}_0$	$\hat{\beta}_1$	$SD(\hat{\beta}_0)$	$SD(\hat{\beta}_1)$	$\hat{\beta}_0$	$\hat{\beta}_1$	$SD(\hat{\beta}_0)$	$SD(\hat{\beta}_1)$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
0	0.979	−0.638	0.089	0.094	0.978	−0.595	0.067	0.054	0.896	0.070
15	0.908	−0.654	0.075	0.100	0.948	−0.598	0.072	0.061	0.598	0.066
25	0.848	−0.671	0.084	0.129	0.902	−0.616	0.067	0.054	0.418	0.073
50	0.754	−0.710	0.072	0.101	0.806	−0.651	0.071	0.061	0.226	0.062

†$E(\xi^2|W)$, $E(\xi^3|W)$ and $E(\xi^4|W)$ are estimated by the non-parametric kernel regression method.

the censoring rate increases. In the latter case, it may be preferable to use the bootstrap method to assess the estimation variability, as suggested in Ma and Yin (2010) and Wang *et al.* (2012). For example, we performed additional bootstraping for the 50% censoring rate case in Table 2. The resulting bootstrap standard deviation is (0.046 0.079 0.122 0.065), which is much closer to the empirical standard deviation (0.062, 0.071, 0.137, 0.071) than the estimated asymptotic standard deviation (0.028, 0.028, 0.102, 0.075).

In the above evaluations, we demonstrate that the RMM2-ISE-method can accurately estimate the covariate effect when the sample size is more than 400 and the censoring rate is less than 50%. Further, RMM2-ISE has better efficiency and smaller mean-squared errors than the IWLS estimator. This encourages us to use the RMM2-ISE-method to analyse the A5175 study data, as we demonstrate in the next section. Moreover, we show that the asymptotic standard deviations are close to the true standard deviations when the observed sample size is sufficient. Finally, we show that the misspecification of $E(\xi_i^3|W_i)$ and $E(\xi_i^4|W_i)$ does not affect the estimations for the parameters β_0 and β_1. Thus, in practice, we can estimate the conditional moments roughly by constructing simple models between W and the power functions of the residuals, such as linear models. This was also justified in Wang *et al.* (2008), which showed that the estimation procedures using the true and the estimated moment functions have similar performance.

Finally, we also perform simulation studies when $E(\xi_i^2|W_i)$ in the IWLS method and $E(\xi_i^3|W_i)$ and $E(\xi_i^4|W_i)$ in the RMM2-ISE-method are estimated by using the non-parametric Nadaraya–Watson kernel method for sample size 800. Compared with IWLS, RMM2-ISE gives less biased results and has smaller variation for estimating the covariate effect. In general, the estimators in Table 3 show larger biases and variations compared with the results in Table 1.

3.3. Analysis of the A5175 study data

We apply the RMM2-ISE-method to the A5175 study data, which aims to evaluate the safety of the antiretroviral regimens. We find that RMM2-ISE gives a better fit to the A5175 study data compared with the commonly used AFT–Weibull–ML method.

We use a total of 1008 patients who have been assigned to open-label antiretroviral therapy with efavirenz plus lamivudine–zidovudine and atazanavir plus didanosine-EC plus emtricitabine treatment arms. A total of 460 patients have their safety events censored, resulting in a censoring rate of 46%. For each patient, we compute the mean of the CD4 cell counts before his or her safety event occurs. To stabilize the numerical computations, we standardize the event times and mean CD4 cell counts by their sample standard deviations, which are approximately 40 and 160 respectively. The transformation is monotone so it does not affect the following inference.

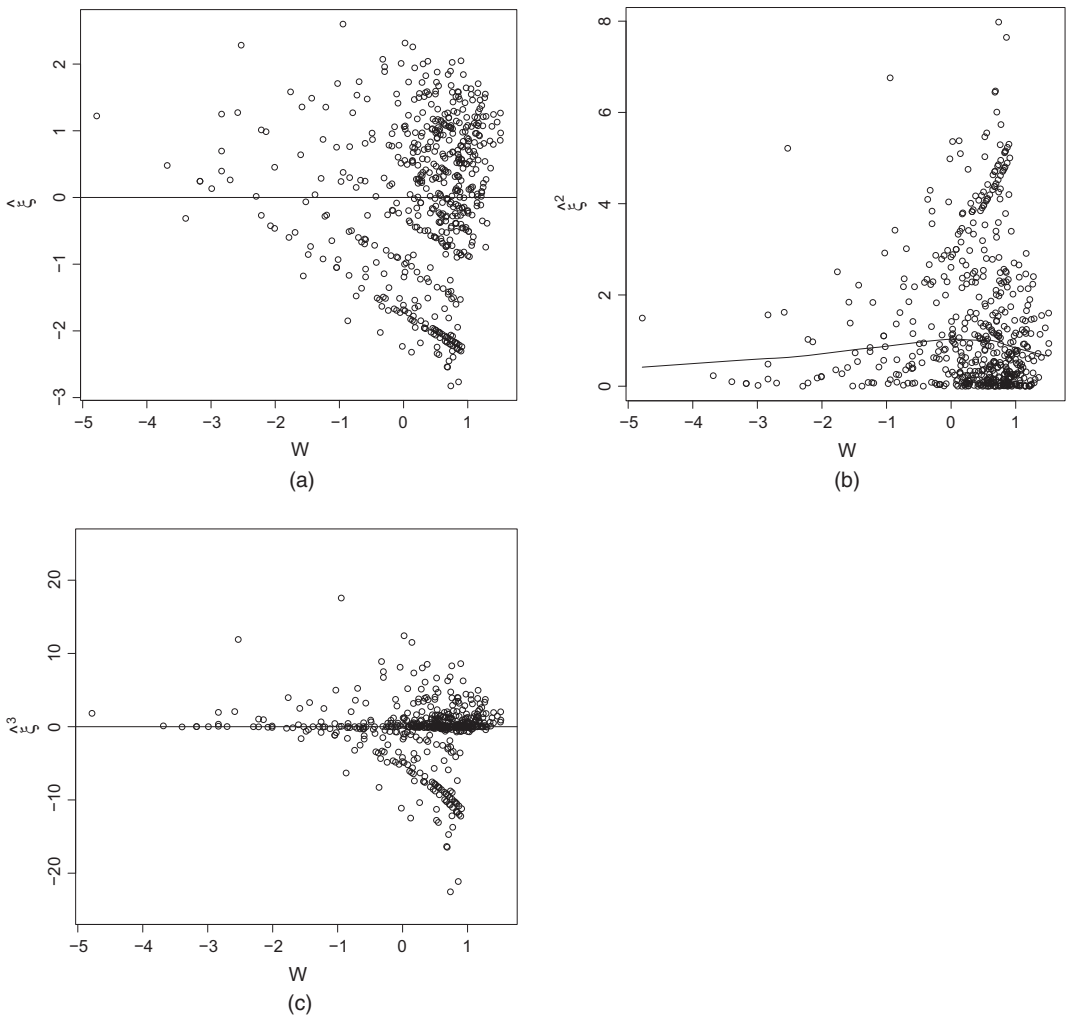

Fig. 1. Preliminary analysis results for the A5175 study data: (a) residual *versus* covariate; (b) residual squared *versus* covariate and a local regression line describing the relationship between ξ^2 and the covariate; (c) scatter plot of the covariate residual cubed and the estimated third moment of the error distribution as a function of the covariate

We denote the standardized event time as T_i and the logarithm of the standardized mean CD4 cell counts as W_i. We first fit the complete data with the linear model

$$\log(T_i) = \beta_0 + \beta_1 W_i + \xi_i,$$

such that $E(\xi_i | W_i) = 0$. Here we use only the non-censored cases to do the initial analysis because our score function in expression (2) is constructed for the non-censored cases only. Further, the data set contains 548 observed survival times, which are sufficient to reveal the general pattern of the error distribution. We plot the residuals $\hat{\xi}_i = \log(T_i) - \hat{\beta}_0 - \hat{\beta}_1 W_i$ *versus* the covariate in Fig. 1(a), where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators of β_0 and β_1 respectively. The residuals are centred at zero, which suggests that the model is adequate to capture the mean structure. Further, the error variation becomes larger when the covariate value increases, which implies

a dependence of the error variance on the covariate. To explore this dependence, we plot the residual squares $\hat{\xi}_i^2$ against the covariates in Fig. 1(b). The plot shows that the variation has a non-linear relationship with W_i. We therefore enrich the linear mean model by further modelling the variance $\sigma^2(W_i, \boldsymbol{\gamma})$. We considered various non-linear forms of $\sigma^2(W_i, \boldsymbol{\gamma})$ and found the form $\sigma^2(W_i, \boldsymbol{\gamma}) = (\gamma_0 + \gamma_1 W_i)^2$ to be both adequate and parsimonious, in that it captures the variability pattern well, and it is simple and yields the smallest estimation variability for $\hat{\boldsymbol{\beta}}$, and this $\hat{\boldsymbol{\beta}}$ is closest to that from the IWLS method among all the non-linear models that we experimented with. Because the misspecification of σ^2 may lead to inconsistent estimators, in practice, we suggest first to use proper variance modelling tools, such as graphical tools, to determine suitable functional forms for $\sigma^2(W_i, \boldsymbol{\gamma})$. After that, we can select the resulting $\hat{\boldsymbol{\beta}}$ from RMM2-ISE which are reasonably close to that from IWLS, because IWLS is always a consistent method regardless of whether the variance form is correctly specified. Finally, we can refine our choices by comparing the variances of $\hat{\boldsymbol{\beta}}$ among the possible candidate variance models.

We implemented RMM2-ISE-estimation on this specific model, and obtained the estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1) = (-0.75, 1.00, 1.25, -0.047)$, with associated standard errors $\{\mathrm{SD}(\hat{\beta}_0), \mathrm{SD}(\hat{\beta}_1), \mathrm{SD}(\hat{\gamma}_0), \mathrm{SD}(\hat{\gamma}_1)\} = (0.047, 0.056, 0.021, 0.049)$. The 95% confidence intervals for the parameters $(\beta_0, \beta_1, \gamma_0, \gamma_1)$ are $\{(-0.84, -0.66), (0.89, 1.11), (1.20, 1.29), (-0.14, 0.049)\}$, which show a significant effect of the CD4 cell counts on the primary safety end point. The covariate effect γ_1 is not significant, which coincides with the local regression line that we added in Fig. 1(b). The local regression technique was proposed by Cleveland (1979). It uses local segments of data to build a function non-parametrically to describe the relationship between the response and the covariate. It can be seen that the local regression line is nearly flat, which suggests that there is no statistically significant effect from the covariate. We also performed IWLS estimation and obtained $(\hat{\beta}_0, \hat{\beta}_1) = (-0.74, 0.99)$, with associated standard errors $\{\mathrm{SD}(\hat{\beta}_0), \mathrm{SD}(\hat{\beta}_1)\} = (0.052, 0.056)$. To obtain the second moment $\sigma^2(W_i)$ as the weight, we first form the regression residuals. Then we propose a working model for $\sigma^2(W_i)$ that is the same as the second-moment model that was used in the RMM2-ISE method, i.e. let $\sigma^2(W_i) = (\gamma_0 + \gamma_1 W_i)^2$, and then perform the usual regression analysis to estimate the parameters in the model and hence obtain the second moment. The results show that RMM2-ISE-estimation is as efficient as the IWLS method. The similar efficiency is not unexpected because, as shown in Fig. 1(c), the estimated conditional third moments of the error terms, i.e. $E(\hat{\xi}_i^3 | W_i)$, are nearly 0. In fact, when we regress $\hat{\xi}_i^3$ on W_i, the resulting intercept is 0.0036 with confidence interval $(-0.02, 0.020)$, and the resulting covariate effect is 0.0004 with confidence interval $(-0.0044, 0.011)$. However, from another aspect, the analysis does demonstrate that the RMM2-ISE-method is at least as efficient as the IWLS method. Therefore we employ the RMM2-ISE-method for the subsequent analyses which ensures that the estimators have variances that are no greater than those resulting from the IWLS method.

To compare the performance of the RMM2-ISE-method with that of the commonly used AFT–Weibull–ML method for the Weibull model, we calculated the mean-squared residuals on the logarithmic scale on the basis of the 548 fully observed samples, and we obtained the values 1.93 for the RMM2-ISE- and 4.69 for the AFT–Weibull–ML method respectively. The comparison based on the observed samples was justified and suggested by Little (1992) in the missingness at random framework, which is the setting in which the non-informative censoring belongs. To avoid overfitting, we performed an additional twofold cross-validation. The cross-validation errors (mean-squared predictive error) for the method proposed and the AFT–Weibull–ML method are 1.89 and 4.23 respectively, indicating that the method proposed outperforms the AFT–Weibull–ML method. The RMM2-ISE-method provides a much better fit to the data

than does the AFT–Weibull–ML method, which also implies that the survival time distribution deviates from Weibull.

After demonstrating the better performance of RMM2-ISE in fitting the A5175 study data, we continue to explore the relationship between CD4 cell counts and the time to primary safety end point in subgroups. We further divided the sample by gender and analysed the CD4 cell counts effects for 479 females and 529 males separately. The estimated $\boldsymbol{\beta}$ in the females group is $(\hat{\beta}_0, \hat{\beta}_1) = (0.14, 0.16)$, and the standard errors $\{\mathrm{SD}(\hat{\beta}_0), \mathrm{SD}(\hat{\beta}_1)\} = (0.14, 0.21)$, which give the confidence intervals $\{(-0.13, 0.41), (-0.25, 0.57)\}$. The estimated $\boldsymbol{\beta}$ in the males group is $(\hat{\beta}_0, \hat{\beta}_1) = (-0.36, 0.31)$, and the standard errors $\{\mathrm{SD}(\hat{\beta}_0), \mathrm{SD}(\hat{\beta}_1)\} = (0.12, 0.14)$, which give the confidence intervals $\{(-0.60, -0.12), (0.04, 0.58)\}$. In the females group, the CD4 cell counts do not have a significant positive effect on the primary safety end points, whereas the effect is significant in the males group. Further, the CD4 cell counts effect is higher in the male patients than in the female patients. It is worth mentioning that, when the AFT–Weibull model is used, no difference between female and male patients can be discovered. In this case, the estimators are $(0.91, 0.32)$ and $(0.96, 0.31)$, the standard deviations are $(0.100, 0.09)$ and $(0.109, 0.112)$ and the 95% confident intervals are $\{(0.69, 1.13), (0.11, 0.55)\}$ and $\{(0.76, 1.16), (0.13, 0.49)\}$ for the females and males respectively. In practice, because the CD4 cell counts are positively related to time to adverse events, we suggest giving the antiretroviral regimens at higher CD4 cell counts level to prevent severe side-effects from the drugs. Further, because the CD4 cell counts effects are different in the two genders, we suggest differentiating the drug scheduling for men and women.

Using the RMM2-ISE-method, we develop a strategy to personalize the drug scheduling based on the A5175 data, where the patients are all enrolled at the beginning of the trial and continuously monitored in the trial, as we now describe. We first define a safety cut-off value regarding the primary safety end point. The drug usage is considered to be safe for a patient if the patient's estimated primary safety end point is later than the cut-off value. A patient's CD4 cell counts are taken at the beginning of the trial (baseline), week 8, week 48, week 72, etc. At a measurement time, say at week 48, we collect the CD4 cell counts information on each patient, and collect his or her primary safety event time or his or her censoring time if either has happened. For a patient who has not experienced the primary safety time and who has not been censored, we use the measurement time as the censoring time. We then use the average observed CD4 cell counts and the event or censoring time to obtain the estimator for the coefficients, i.e. $\boldsymbol{\beta}$. Then, for any patient who has not experienced the primary safety event at the 48th week, we use the estimate $\hat{\boldsymbol{\beta}}$ and his or her average observed CD4 cell counts to predict his or her primary safety event time. If the predicted primary safety event time is to the right of the safety cut-off value, the treatment is considered safe for the patient. This patient is eliminated from the current trial and moves to the next treatment phase. We perform this estimation and prediction procedure at weeks 8, 48 and 72 and make corresponding decisions at each measurement time based on the remaining patients in the trial.

We use the 75% sample quantile of the standardized primary safety end points, 2 (corresponding to 79.14 in the original data), as a sample cut-off value. In practice, different and possibly more meaningful cut-off values can be chosen based on existing medical knowledge. We choose to start to treat a patient with the antiretroviral regimens when the lower bound of the estimated confidence interval for the mean of $\log(T_i)$, i.e. $\hat{\beta}_0 + \hat{\beta}_1 W_i - 1.96\{(1, W_i)^{\mathrm{T}} \hat{\Sigma} (1, W_i)\}^{1/2}$, is greater than $\log(2)$, where $\hat{\Sigma}$ is the estimated variance–covariance matrix for $\hat{\boldsymbol{\beta}}$. We perform the analysis in the following three groups of patients. Group 1 contains patients who have only baseline CD4 cell counts recorded. Group 2 contains patients who have CD4 cell counts measured at and before the 48th week. Group 3 contains patients who have CD4 cell counts measured at

and before the 96th week. The results show that, in group 1, 95 out of 188 (50.5%) patients have the lower bound of the estimated confidence interval smaller than log(2). Further, in group 2 and group 3, the ratios are 132 out of 201 (66%) and 94 out of 131 (72%) respectively. Therefore, in these three groups, we can start to treat 50.5%, 66% and 72% of the patients at the baseline randomization time, the 48th week or the 96th week respectively. Since the CD4 cell counts are obtained before the primary safety end points, the strategy allows the patients to be treated earlier when the evidence of the treatment safety is sufficient and thus improves the efficiency of delivering the safe treatments to the patients.

In conclusion, we compared the RMM2-ISE-method with the AFT–Weibull–ML method. The RMM2-ISE- outperforms the AFT–Weibull–ML method in giving smaller fitted mean-squared residuals. Further, we discovered that the positive CD4 cell counts effects in men are higher than that in women on average, whereas this pattern is not captured by the AFT–Weibull–ML method. Finally, we propose a strategy for personalizing drug scheduling based on the mean of the repeatedly measured CD4 cell counts. The strategy allows early treatment delivery to patients based on their CD4 cell counts information, and ultimately enhances the treatment efficiency.

4. Discussion

This work was motivated by the A5175 study (Campbell *et al.*, 2012). We intend to use the short-term CD4 cell counts to infer the primary safety end points. The complex data configuration motivates us to construct a second-order restricted moment model which models the additional variance structures that are observed from the data. We propose a non-parametric imputed version of the semiparametric efficient method for parameter estimations to handle censoring. The theoretical derivations show that the resulting estimators are consistent and asymptotically normally distributed. The efficiency of RMM2-ISE is demonstrated to be better than that of IWLS estimators. When fitting the A5175 study data, the RMM2-ISE-method outperforms the AFT–Weibull–ML method in terms of having smaller mean-squared residuals.

In the A5175 data analysis, because of the limitation of the univariate kernel specification, we did not include multiple covariates in the regression function. The method can be extended to include multiple covariates through utilizing multivariate kernels. Such an extension will enhance the applicability of the model in more general situations.

In conclusion, to analyse the A5175 data, the second-order restricted moment model avoids the model assumptions on the full likelihood and is more flexible than parametric models. Further, in terms of parameter estimation, the RMM2-ISE-method takes advantage of the additional information in the variance structure and has better efficiency than the IWLS method. In general, the RMM2-ISE-approach provides a more robust and efficient way of analysing post-trial data.

We have assumed that the censoring process is independent of the covariates and the survival process for simplicity. This assumption can be relaxed to allow the censoring time to depend on the covariates w_j. In this case, we can use a non-parametric kernel-based Kaplan–Meier estimator

$$\hat{G}(t_j|W_j = w_j) = \prod_{x_i \leqslant t_j} \left\{ 1 - \frac{(1-\Delta_i) K_h(w_i - w_j)}{\sum_{k=1}^{n} I(x_k \geqslant x_i) K_h(w_k - w_j)} \right\}$$

in equation (4). However, the subsequent development will also need to be adapted to reflect the covariate-dependent nature of the censoring process and the analysis will be more complex.

References

Campbell, T. B., Smeaton, L. M., Kumarasamy, N., Flanigan, T., Klingman, K. L., Firnhaber, C., Grinsztejn, B., Hosseinipour, M. C., Kumwenda, J., Lalloo, U., Riviere, C., Sanchez, J., Melo, M., Supparatpinyo, K., Tripathy, S., Martinez, A. I., Nair, A., Walawander, A., Moran, L., Chen, Y., Snowden, W., Rooney, J. F., Uy, J., Schooley, R. T., De Gruttola, V., Hakim, J. G. and the Study Team of the ACTG (2012) Efficacy and safety of three antiretroviral regimens for initial treatment of hiv-1: a randomized clinical trial in diverse multinational settings. *PLOS Med*., **9**, no. 8, article e1001290.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Ass.*, **74**, 829–836.

Devroye, L. (1981) On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.*, **9**, 1310–1319.

Gallant, A. R. (2009) *Nonlinear Statistical Models*. Hoboken: Wiley.

Hirsch, M. S. (2008) Initiating therapy: when to start, what to use. *J. Infect. Dis.*, **197**, suppl. 3, S252–S260.

Kim, M. and Ma, Y. (2012) The efficiency of the second-order nonlinear least squares estimator and its extension. *Ann. Inst. Statist. Math.*, **64**, 1–14.

Klein, J. P. and Moeschberger, M. L. (2010) *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999) A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Am. Statist. Ass.*, **94**, 1147–1160.

Little, R. J. (1992) Regression with missing x's: a review. *J. Am. Statist. Ass.*, **87**, 1227–1237.

Ma, Y. and Yin, G. (2010) Semiparametric median residual life model and inference. *Can. J. Statist.*, **38**, 665–679.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Wang, Y., Garcia, T. P. and Ma, Y. (2012) Nonparametric estimation for censored mixture data with application to the cooperative Huntingtons observational research trial. *J. Am. Statist Ass.*, **107**, 1324–1338.

Wang, S., Joshi, S., Mboudjeka, I., Liu, F., Ling, T., Goguen, J. D. and Lu, S. (2008) Relative immunogenicity and protection potential of candidate yersinia pestis antigens against lethal mucosal plague challenge in balb/c mice. *Vaccine*, **26**, 1664–1674.

Wang, L. and Leblanc, A. (2008) Second-order nonlinear least squares estimation. *Ann. Int. Statist. Math.*, **60**, 883–900.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'A second order semiparametric method for survival analysis, with application to an AIDS clinical trial study: Appendix'.