J. R. Statist. Soc. B (2007)
69, *Part* 4, *pp.* 507–564

Maximum likelihood estimation in semiparametric regression models with censored data

D. Zeng and D. Y. Lin

University of North Carolina, Chapel Hill, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, January 31st, 2007*, Professor T. J. Sweeting *in the Chair*]

Summary. Semiparametric regression models play a central role in formulating the effects of covariates on potentially censored failure times and in the joint modelling of incomplete repeated measures and failure times in longitudinal studies. The presence of infinite dimensional parameters poses considerable theoretical and computational challenges in the statistical analysis of such models. We present several classes of semiparametric regression models, which extend the existing models in important directions. We construct appropriate likelihood functions involving both finite dimensional and infinite dimensional parameters. The maximum likelihood estimators are consistent and asymptotically normal with efficient variances. We develop simple and stable numerical techniques to implement the corresponding inference procedures. Extensive simulation experiments demonstrate that the inferential and computational methods proposed perform well in practical settings. Applications to three medical studies yield important new insights. We conclude that there is no reason, theoretical or numerical, not to use maximum likelihood estimation for semiparametric regression models. We discuss several areas that need further research.

Keywords: Counting process; EM algorithm; Generalized linear mixed models; Joint models; Multivariate failure times; Non-parametric likelihood; Profile likelihood; Proportional hazards; Random effects; Repeated measures; Semiparametric efficiency; Survival data; Transformation models

1. Introduction

The Cox (1972) proportional hazards model is the corner-stone of modern survival analysis. The model specifies that the hazard function of the failure time conditional on a set of possibly time varying covariates is the product of an arbitrary base-line hazard function and a regression function of the covariates. Cox (1972, 1975) introduced the ingenious partial likelihood principle to eliminate the infinite dimensional base-line hazard function from the estimation of regression parameters with censored data. In a seminal paper, Andersen and Gill (1982) extended the Cox regression model to general counting processes and established the asymptotic properties of the maximum partial likelihood estimator and the associated Breslow (1972) estimator of the cumulative base-line hazard function via the elegant counting process martingale theory. The maximum partial likelihood estimator and the Breslow estimator can be viewed as non-parametric maximum likelihood estimators (NPMLEs) in that they maximize the non-parametric likelihood in which the cumulative base-line hazard function is regarded as an infinite dimensional parameter (Andersen *et al.* (1993), pages 221–229 and 481–483, and Kalbfleisch and Prentice (2002), pages 114–128).

Address for correspondence: D. Y. Lin, Department of Biostatistics, CB 7420, University of North Carolina, Chapel Hill, NC 27599-7420, USA.
E-mail: lin@bios.unc.edu

1369–7412/07/69507

The proportional hazards assumption is often violated in scientific studies, and other semiparametric models may provide more accurate or more concise summarization of data. Under the proportional odds model (Bennett, 1983), for instance, the hazard ratio between two sets of covariate values converges to 1, rather than staying constant, as time increases. The NPMLE for this model was studied by Murphy *et al.* (1997). Both the proportional hazards and the proportional odds models belong to the class of linear transformation models which relates an unknown transformation of the failure time linearly to covariates (Kalbfleisch and Prentice (2002), page 241). Dabrowska and Doksum (1988), Cheng *et al.* (1995) and Chen *et al.* (2002) proposed general estimators for this class of models, none of which are asymptotically efficient. The class of linear transformation models is confined to traditional survival (i.e. single-event) data and time invariant covariates.

As an example of non-proportional hazards structures, Fig. 1 displays (in the full curves) the Kaplan–Meier estimates of survival probabilities for the chemotherapy and chemotherapy plus radiotherapy groups of gastric cancer patients in a randomized clinical trial (Stablein and Koutrouvelis, 1985). The crossing of the two survival curves is a strong indication of crossing hazards. This is common in clinical trials because the patients who receive the more aggressive intervention (e.g. radiotherapy or transplantation) are at elevated risks of death initially but may enjoy considerable long-term survival benefits if they can tolerate the intervention. Crossing hazards cannot be captured by linear transformation models. The use of the proportional hazards model could yield very misleading results in such situations.

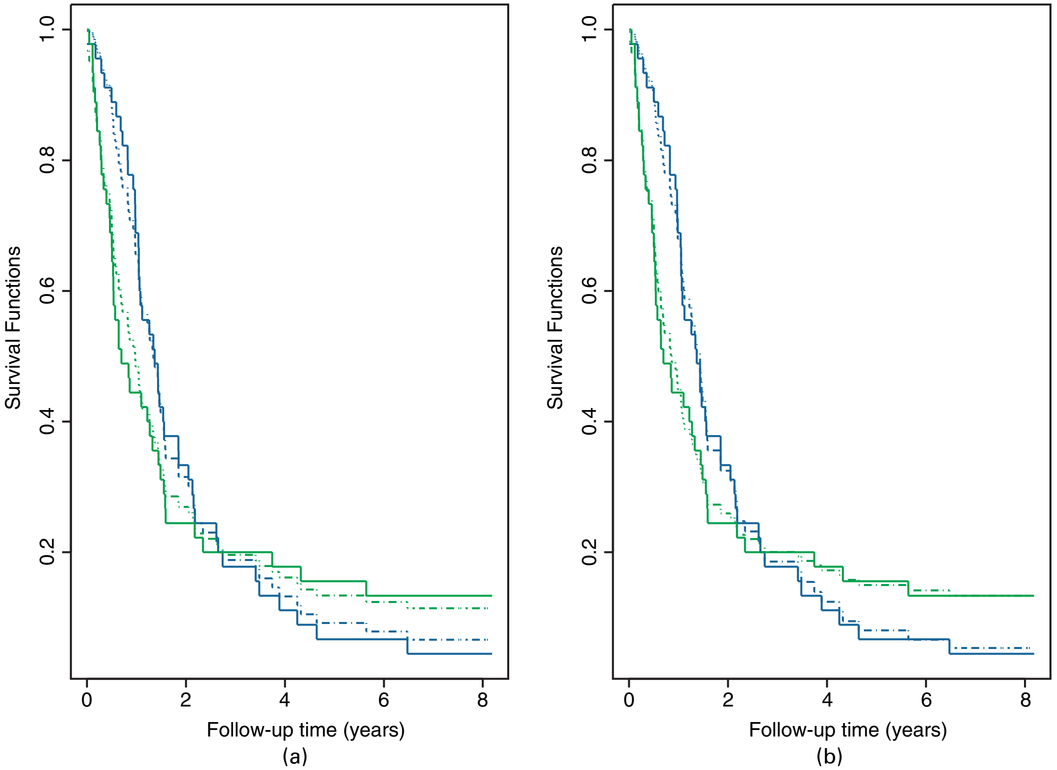

Fig. 1. Kaplan–Meier (———) and model-based estimates (- · - · - ·) of survival functions for gastrointestinal tumour patients (the chemotherapy and combined therapy patients are indicated by blue and green respectively): (a) model (3); (b) model (4)

Multivariate or dependent failure time data arise when each study subject can potentially experience several events or when subjects are sampled in clusters (Kalbfleisch and Prentice (2002), chapters 8–10). It is natural and convenient to represent the dependence of related failure times through frailty or random effects (Clayton and Cuzick, 1985; Oakes, 1989, 1991; Hougaard, 2000). The NPMLE of the proportional hazards model with gamma frailty was studied by Nielsen *et al.* (1992), Klein (1992), Murphy (1994, 1995), Andersen *et al.* (1997) and Parner (1998). Gamma frailty induces a very restrictive form of dependence, and the proportional hazards assumption fails more often with complex multivariate failure time data than with univariate data. The focus of the existing literature on the proportional hazards gamma frailty model is due to its mathematical tractability. Cai *et al.* (2002) proposed estimating equations for linear transformation models with random effects for clustered failure time data. Zeng *et al.* (2005) studied the NPMLE for the proportional odds model with normal random effects and found the estimators of Cai *et al.* (2002) to be considerably less efficient.

Lin (1994) described a colon cancer study in which the investigators wished to assess the efficacy of adjuvant therapy on recurrence of cancer and death for patients with resected colon cancer. By characterizing the dependence between recurrence of cancer and death through a random effect, one could properly account for the informative censoring caused by death on recurrence of cancer and accurately predict a patient’s survival outcome given his or her cancer recurrence time. However, random-effects models for multiple types of events have received little attention in the literature.

In longitudinal studies, data are often collected on repeated measures of a response variable as well as on the time to the occurrence of a certain event. There is a tremendous recent interest in joint modelling, in which models for the repeated measures and failure time are assumed to depend on a common set of random effects. Such models can be used to assess the joint effects of base-line covariates (such as treatments) on the two types of outcomes, to study the effects of potentially mismeasured time varying covariates on the failure time and to adjust for informative drop-out in the analysis of repeated measures. The existing literature (e.g. Wulfsohn and Tsiatis (1997), Hogan and Laird (1997) and Henderson *et al.* (2000)) has been focused on the linear mixed model for repeated measures and the proportional hazards model with normal random effects for the failure time.

The linear mixed model is confined to continuous repeated measures with normal error. In addition, the transformation of the response variable is assumed to be known. Inference under random-effects models is highly non-robust to misspecification of transformation. Our experience in human immunodeficiency virus (HIV) and acquired immune deficiency syndrome research shows that different transformations of CD cell counts often yield conflicting results. Thus, it would be desirable to employ semiparametric models (e.g. linear transformation models) for continuous repeated measures, so that a parametric specification of the transformation or distribution can be avoided. This kind of model has not been studied even without the task of joint modelling, although econometricians (Horowitz (1998), chapter 5) have proposed inefficient estimators for univariate responses.

As evident from the above description, the existing semiparametric regression models, although very useful, have important limitations and, in most cases, lack efficient estimators or careful theoretical treatments. In this paper, we unify and extend the current literature, providing a comprehensive methodology with strong theoretical underpinning. We propose a very general class of transformation models for counting processes which encompasses linear transformation models and which accommodates crossing hazards, time varying covariates and recurrent events. We then extend this class of models to dependent failure time data (including recurrent events, multiple types of events and clustered failure time data) by incorporating a rich family of multivariate random effects. Furthermore, we present a broad class of joint models by specifying random-effects transformation models for the failure time and generalized linear mixed models for (discrete or continuous) repeated measures. We also propose a semiparametric linear mixed model for continuous repeated measures, under which the transformation of the response variable is completely unspecified.

We establish the consistency, asymptotic normality and asymptotic efficiency of the NPMLEs for the proposed models by appealing to modern empirical process theory (van der Vaart and Wellner, 1996) and semiparametric efficiency theory (Bickel *et al.*, 1993). In fact, we develop a very general asymptotic theory for non-parametric maximum likelihood estimation with censored data. Our general theory can be used to derive asymptotic results for many existing semiparametric models which are not covered in this paper as well as those to be invented in the future. Simulation studies show that the asymptotic approximations are accurate for practical sample sizes.

It is widely believed that NPMLEs are intractable computationally. This perception has motivated the development of *ad hoc* estimators which are less efficient statistically. We present in this paper simple and effective methods to calculate the NPMLEs and to implement the corresponding inference procedures. These methods apply to a wide variety of semiparametric models with censored data and make the NPMLEs computationally more feasible than the *ad hoc* estimators (when the latter exist). Their usefulness is amply demonstrated through simulated and real data.

As hinted in the discussion thus far, we are suggesting the following strategies in the research and practice of survival analysis and related fields.

(a) Use the new class of transformation models to analyse failure time data.
(b) Make routine use of random-effects models for multivariate failure time data.
(c) Choose normal random effects over gamma frailty.
(d) Determine transformations of continuous response variables non-parametrically.
(e) Formulate multiple types of outcome measures with semiparametric joint models.
(f) Adopt maximum likelihood estimation for semiparametric regression models.
(g) Rely on modern empirical process theory as the primary mathematical tool.

We shall elaborate on these points in what follows, particularly at the end. In addition, we shall pose a wide range of open problems and outline several directions for future research.

2. Semiparametric models

2.1. Transformation models for counting processes

The class of linear transformation models relates an unknown transformation of the failure time T linearly to a vector of (time invariant) covariates Z:

$$H(T) = -\beta^{\mathrm{T}} Z + \varepsilon, \tag{1}$$

where $H(\cdot)$ is an unspecified increasing function, β is a set of unknown regression parameters and ε is a random error with a parametric distribution. The choices of the extreme value and standard logistic error distributions yield the proportional hazards and proportional odds models respectively.

Remark 1. The familiar linear model form of equation (1) is very appealing. Since the transformation $H(\cdot)$ is arbitrary, the parametric assumption on ε should not be viewed as restrictive. In fact, without Z, there is always a transformation such that ε has any given distribution.

We extend equation (1) to allow time varying covariates and recurrent events. Let $N^*(t)$ be the counting process recording the number of events that have occurred by time t, and let $Z(\cdot)$ be a vector of possibly time varying covariates. We specify that the cumulative intensity function for $N^*(t)$ conditional on $\{Z(s); s \leqslant t\}$ takes the form

$$\Lambda(t|Z) = G\left[\int_0^t R^*(s) \exp\{\beta^{\mathrm{T}} Z(s)\} \,\mathrm{d}\Lambda(s)\right], \tag{2}$$

where G is a continuously differentiable and strictly increasing function, $R^*(\cdot)$ is an indicator process, β is a vector of unknown regression parameters and $\Lambda(\cdot)$ is an unspecified increasing function. For survival data, $R^*(t) = I(T \geqslant t)$, where $I(\cdot)$ is the indicator function; for recurrent events, $R^*(\cdot) = 1$. It is useful to consider the class of Box–Cox transformations

$$G(x) = \frac{(1+x)^\rho - 1}{\rho}, \qquad \rho \geqslant 0,$$

with $\rho = 0$ corresponding to $G(x) = \log(1+x)$ and the class of logarithmic transformations

$$G(x) = \frac{\log(1+rx)}{r}, \qquad r \geqslant 0,$$

with $r = 0$ corresponding to $G(x) = x$. The choice of $G(x) = x$ yields the familiar proportional hazards or intensity model (Cox, 1972; Andersen and Gill, 1982). If $N^*(\cdot)$ has a single jump at the survival time T and Z is time invariant, then equation (2) reduces to equation (1).

Remark 2. Specifying the function G while leaving the function Λ unspecified is equivalent to specifying the distribution of ε while leaving the function H unspecified. Non-identifiability arises if both G and Λ (or both H and ε) are unspecified and $\beta = 0$; see Horowitz (1998), page 169.

To capture the phenomenon of crossing hazards as seen in Fig. 1, we consider the heteroscedastic version of linear transformation models

$$H(T) = -\beta^{\mathrm{T}} Z + \exp(-\gamma^{\mathrm{T}} \tilde{Z})\varepsilon,$$

where $\tilde{Z}$ is a set of (time invariant) covariates and γ is the corresponding vector of regression parameters. For notational simplicity, we assume that $\tilde{Z}$ is a subset of Z, although this assumption is not necessary. Under this formulation, the hazard functions that are associated with different values of $\tilde{Z}$ can cross and the hazard ratio can invert over time. To accommodate such scenarios as well as recurrent events and time varying covariates, we extend equation (2) as follows:

$$\Lambda(t|Z) = G\left(\left[\int_0^t R^*(s) \exp\{\beta^{\mathrm{T}} Z(s)\} \,\mathrm{d}\Lambda(s)\right]^{\exp(\gamma^{\mathrm{T}} \tilde{Z})}\right). \tag{3}$$

For survival data, model (3) with $G(x) = x$ is similar to the heteroscedastic hazard model of Hsieh (2001), who proposed to fit his model by the method of histogram sieves.

Under model (3) and Hsieh's model, the hazard function is infinite at time 0 if $\gamma^{\mathrm{T}} \tilde{Z} < 0$. This feature causes some technical difficulty. Thus, we propose the following modification:

$$\Lambda(t|Z) = G\left(\left[1 + \int_0^t R^*(s) \exp\{\beta^{\mathrm{T}} Z(s)\} \,\mathrm{d}\Lambda(s)\right]^{\exp(\gamma^{\mathrm{T}} \tilde{Z})}\right) - G(1). \tag{4}$$

If $\gamma=0$, equation (4) reduces to equation (2) by redefining $G(1+x)-G(1)$ as $G(x)$. For survival data, the conditional hazard function under model (4) with $G(x)=x$ becomes

$$\exp\{\beta^{\mathrm{T}} Z(s)+\gamma^{\mathrm{T}}\tilde{Z}\}\left[1+\int_0^t \exp\{\beta^{\mathrm{T}} Z(s)\}\,\mathrm{d}\Lambda(s)\right]^{\exp(\gamma^{\mathrm{T}}\tilde{Z})-1}\lambda(t),$$

where $\lambda(t)=\Lambda'(t)$. Here and in what follows $g'(x)=\mathrm{d}g(x)/\mathrm{d}x$. This model is similar to the cross-effects model of Bagdonavicius *et al.* (2004), who fitted their model by modifying the partial likelihood.

Let C denote the censoring time, which is assumed to be independent of $N^*(\cdot)$ conditional on $Z(\cdot)$. For a random sample of n subjects, the data consist of $\{N_i(t), R_i(t), Z_i(t); t\in[0,\tau]\}$ $(i=1,\ldots,n)$, where $R_i(t)=I(C_i\geqslant t)\,R_i^*(t)$, $N_i(t)=N_i^*(t\wedge C_i)$, $a\wedge b=\min(a,b)$ and τ is the duration of the study. For general censoring and truncation patterns, we define $N_i(t)$ as the number of events that are observed by time t on the ith subject, and $R_i(t)$ as the indicator on whether the ith subject is at risk at t.

Write $\lambda(t|Z)=\Lambda'(t|Z)$ and $\theta=(\beta^{\mathrm{T}},\gamma^{\mathrm{T}})^{\mathrm{T}}$. Assume that censoring is non-informative about the parameters θ and $\Lambda(\cdot)$. Then the likelihood for θ and $\Lambda(\cdot)$ is proportional to

$$\prod_{i=1}^{n}\prod_{t\leqslant\tau}\{R_i(t)\,\lambda(t|Z_i)\}^{\mathrm{d}N_i(t)}\exp\left\{-\int_0^{\tau} R_i(t)\,\lambda(t|Z_i)\,\mathrm{d}t\right\},\qquad(5)$$

where $\mathrm{d}N_i(t)$ is the increment of N_i over $[t,t+\mathrm{d}t)$.

2.2. Transformation models with random effects for dependent failure times

For recurrent events, models (2)–(4) assume that the occurrence of a future event is independent of the prior event history unless such dependence is represented by suitable time varying covariates. It is inappropriate to use such time varying covariates in randomized clinical trials because the inclusion of a post-randomization response variable in the model will attenuate the estimator of treatment effect. It is more appealing to characterize the dependence of recurrent events through random effects or frailty. Frailty is also useful in formulating the dependence of several types of events on the same subject or the dependence of failure times among individuals of the same cluster. To accommodate all these types of data structure, we represent the underlying counting processes by $N_{ikl}^*(\cdot)$ $(i=1,\ldots,n;\ k=1,\ldots,K;\ l=1,\ldots,n_{ik})$, where i pertains to a subject or cluster, k to the type of event and l to individuals within a cluster; see Andersen *et al.* (1993), pages 660–662. The specific choices of $K=n_{ik}=1$, $n_{ik}=1$ and $K=1$ correspond to recurrent events, multiple types of events and clustered failure times respectively. For the colon cancer study that was mentioned in Section 1, $K=2$ (and $n_{ik}=1$), with $k=1$ and $k=2$ representing cancer recurrence and death.

The existing literature is largely confined to proportional hazards or intensity models with gamma frailty, under which the intensity function for $N_{ikl}^*(t)$ conditional on covariates $Z_{ikl}(t)$ and frailty ξ_i takes the form

$$\lambda_k(t|Z_{ikl};\xi_i)=\xi_i\,R_{ikl}^*(t)\,\exp\{\beta^{\mathrm{T}} Z_{ikl}(t)\}\,\lambda_k(t),\qquad(6)$$

where ξ_i $(i=1,\ldots,n)$ are gamma-distributed random variables, R_{ikl}^* is analogous to R_i^* and $\lambda_k(\cdot)$ $(k=1,\ldots,K)$ are arbitrary base-line functions. Murphy (1994, 1995) and Parner (1998) established the asymptotic theory of the NPMLEs for recurrent events without covariates and for clustered failure times with covariates respectively.

Remark 3. Kosorok *et al.* (2004) studied the proportional hazards frailty model for univariate survival data. The induced marginal model (after integrating out the frailty) is a linear transformation model in the form of equation (1).

We assume that the cumulative intensity function for $N^*_{ikl}(t)$ takes the form

$$\Lambda_k(t|Z_{ikl};b_i)=G_k\left[\int_0^t R^*_{ikl}(s)\exp\{\beta^{\mathrm{T}}Z_{ikl}(s)+b_i^{\mathrm{T}}\tilde{Z}_{ikl}(s)\}\,\mathrm{d}\Lambda_k(s)\right], \tag{7}$$

where G_k $(k=1,\ldots,K)$ are analogous to G of Section 2.1, $\tilde{Z}_{ikl}$ is a subset of Z_{ikl} plus the unit component, b_i $(i=1,\ldots,n)$ are independent random vectors with multivariate density function $f(b;\gamma)$ indexed by a set of parameters γ and $\Lambda_k(\cdot)$ $(k=1,\ldots,K)$ are arbitrary increasing functions. Equation (7) is much more general than equation (6) in that it accommodates non-proportional hazards or intensity models and multiple random effects that may not be gamma distributed. It is particularly appealing to allow normal random effects, which, unlike gamma frailty, have unrestricted covariance matrices. In light of the linear transformation model representation, normal random effects are more natural than gamma frailty, even for the proportional hazards model. Computationally, normal distributions are more tractable than others, especially for high dimensional random effects.

Write $\theta=(\beta^{\mathrm{T}},\gamma^{\mathrm{T}})^{\mathrm{T}}$. Let C_{ikl}, $N_{ikl}(\cdot)$ and $R_{ikl}(\cdot)$ be defined analogously to C_i, $N_i(\cdot)$ and $R_i(\cdot)$ of Section 2.1. Assume that C_{ikl} is independent of $N^*_{ikl}(\cdot)$ and b_i conditional on $Z_{ikl}(\cdot)$ and non-informative about θ and Λ_k $(k=1,\ldots,K)$. The likelihood for θ and Λ_k $(k=1,\ldots,K)$ is

$$\begin{aligned}\prod_{i=1}^n\int_b\prod_{k=1}^K\prod_{l=1}^{n_{ik}}\prod_{t\leqslant\tau}\Bigg(&R_{ikl}(t)\,\lambda_k(t)\exp\{\beta^{\mathrm{T}}Z_{ikl}(t)+b^{\mathrm{T}}\tilde{Z}_{ikl}(t)\}\\&\times G'_k\left[\int_0^t R_{ikl}(s)\exp\{\beta^{\mathrm{T}}Z_{ikl}(s)+b^{\mathrm{T}}\tilde{Z}_{ikl}(s)\}\,\mathrm{d}\Lambda_k(s)\right]\Bigg)^{\mathrm{d}N_{ikl}(t)}\\&\times\exp\left(-G_k\left[\int_0^\tau R_{ikl}(t)\exp\{\beta^{\mathrm{T}}Z_{ikl}(t)+b^{\mathrm{T}}\tilde{Z}_{ikl}(t)\}\,\mathrm{d}\Lambda_k(t)\right]\right)f(b;\gamma)\,\mathrm{d}b,\end{aligned} \tag{8}$$

where $\lambda_k(t)=\Lambda'_k(t)$ $(k=1,\ldots,K)$.

2.3. *Joint models for repeated measures and failure times*

Let Y_{ij} represent a response variable and X_{ij} a vector of covariates that are observed at time t_{ij}, for observation $j=1,\ldots,n_i$ on subject $i=1,\ldots,n$. We formulate these repeated measures through generalized linear mixed models (Diggle *et al.* (2002), section 7.2). The random effects b_i $(i=1,\ldots,n)$ are independent zero-mean random vectors with multivariate density function $f(b;\gamma)$ indexed by a set of parameters γ. Given b_i, the responses $Y_{i1},\ldots,Y_{in_i}$ are independent and follow a generalized linear model with density $f_y(y|X_{ij};b_i)$. The conditional means satisfy

$$g\{E(Y_{ij}|X_{ij};b_i)\}=\alpha^{\mathrm{T}}X_{ij}+b_i^{\mathrm{T}}\tilde{X}_{ij},$$

where g is a known link function, α is a set of regression parameters and $\tilde{X}$ is a subset of X.

As in Section 2.1, let $N^*_i(t)$ denote the number of events which the ith subject has experienced by time t and $Z_i(\cdot)$ be a vector of covariates. We allow $N^*_i(\cdot)$ to take multiple jumps to accommodate recurrent events. If we are interested in adjusting for informative drop-out in the repeated measures analysis, however, $N^*_i(\cdot)$ will take a single jump at the drop-out time. To account for the correlation between $N^*_i(\cdot)$ and the Y_{ij}, we incorporate the random effects b_i into equation (2),

$$\Lambda(t|Z_i;b_i)=G\left[\int_0^t R_i^*(s)\exp\{\beta^{\mathrm{T}} Z_i(s)+(\psi\circ b_i)^{\mathrm{T}}\,\tilde{Z}_i(s)\}\,\mathrm{d}\Lambda(s)\right],$$

where $\tilde{Z}_i$ is a subset of Z_i plus the unit component, ψ is a vector of unknown constants and $v_1\circ v_2$ is the componentwise product of two vectors v_1 and v_2. Typically but not necessarily, $X_{ij}=Z_i(t_{ij})$. It is assumed that $N_i^*(\cdot)$ and the Y_{ij} are independent given b_i, Z_i and X_{ij}.

Write $\theta=(\alpha^{\mathrm{T}},\beta^{\mathrm{T}},\gamma^{\mathrm{T}},\psi^{\mathrm{T}})^{\mathrm{T}}$. Assume that censoring and measurement times are non-informative (Tsiatis and Davidian, 2004). Then the likelihood for θ and $\Lambda(\cdot)$ can be written as

$$\prod_{i=1}^{n}\int_b\prod_{t\leqslant\tau}\{R_i(t)\,\lambda(t|Z_i;b)\}^{\mathrm{d}N_i(t)}\,\exp\left\{-\int_0^{\tau}R_i(t)\,\lambda(t|Z_i;b)\,\mathrm{d}t\right\}\prod_{j=1}^{n_i}f_y(Y_{ij}|X_{ij};b)\,f(b;\gamma)\,\mathrm{d}b, \tag{9}$$

where $\lambda(t|Z;b)=\Lambda'(t|Z;b)$.

It is customary to use the linear mixed model for continuous repeated measures. The normality that is required by the linear mixed model may not hold. A simple strategy to achieve approximate normality is to apply a parametric transformation to the response variable. It is difficult to find the correct transformation in practice, especially when there are outlying observations. As mentioned in Section 1, our experience in analysing HIV data shows that different transformations (such as logarithmic *versus* square root) of CD cell counts or viral loads often lead to conflicting results. Thus, we propose the semiparametric linear mixed model or random-effects linear transformation model

$$\tilde{H}(Y_{ij})=\alpha^{\mathrm{T}}X_{ij}+b_i^{\mathrm{T}}\tilde{X}_{ij}+\varepsilon_{ij}, \tag{10}$$

where $\tilde{H}$ is an unknown increasing function and ε_{ij} $(i=1,\ldots,n;\,j=1,\ldots,n_{ij})$ are independent errors with density function f_ε. If the transformation function $\tilde{H}$ were specified, then equation (10) would reduce to the conventional (parametric) linear mixed model. Leaving the form of $\tilde{H}$ unspecified is in line with the semiparametric feature of the transformation models for event times. There is no intercept in α since it can be absorbed in $\tilde{H}$. Write $\tilde{\Lambda}(y)=\exp\{\tilde{H}(y)\}$. The likelihood for θ, Λ and $\tilde{\Lambda}$ is

$$\prod_{i=1}^{n}\int_b\prod_{t\leqslant\tau}\{R_i(t)\,\lambda(t|Z_i;b)\}^{\mathrm{d}N_i(t)}\exp\left\{-\int_0^{\tau}R_i(t)\,\lambda(t|Z_i;b)\,\mathrm{d}t\right\}$$
$$\times\prod_{j=1}^{n_i}f_\varepsilon[\log\{\tilde{\Lambda}(Y_{ij})\}-\alpha^{\mathrm{T}}X_{ij}-b_i^{\mathrm{T}}\tilde{X}_{ij}]\frac{\tilde{\lambda}(Y_{ij})}{\tilde{\Lambda}(Y_{ij})}f(b;\gamma)\,\mathrm{d}b, \tag{11}$$

where $\tilde{\lambda}(y)=\tilde{\Lambda}'(y)$.

3. Maximum likelihood estimation

The likelihood functions that are given in expressions (5), (8), (9) and (11) can all be written in a generic form

$$L_n(\theta,\mathcal{A})=\prod_{i=1}^{n}\prod_{k=1}^{K}\prod_{l=1}^{n_{ik}}\prod_{t\leqslant\tau}\lambda_k(t)^{\mathrm{d}N_{ikl}(t)}\,\Psi(\mathcal{O}_i;\theta,\mathcal{A}), \tag{12}$$

where $\mathcal{A}=(\Lambda_1,\ldots,\Lambda_K)$, $\mathcal{O}_i$ is the observation on the ith subject or cluster and Ψ is a functional of random process $\mathcal{O}_i$, infinite dimensional parameter $\mathcal{A}$ and d-dimensional parameter θ; expression (11) can be viewed as a special case of expression (8) with $K=2$, $\mathcal{A}=(\Lambda,\tilde{\Lambda})$, $n_{i1}=1$ and $n_{i2}=n_i$, where repeated measures correspond to the second type of failure. To obtain the

Kiefer–Wolfowitz NPMLEs of θ and $\mathcal{A}$, we treat $\mathcal{A}$ as right continuous and replace $\lambda_k(t)$ by the jump size of Λ_k at t, which is denoted by $\Lambda_k\{t\}$. Under model (2) with $G(x) = x$, the NPMLEs are identical to the maximum partial likelihood estimator of β and the Breslow estimator of Λ.

The calculation of the NPMLEs is tantamount to maximizing $L_n(\theta, \mathcal{A})$ with respect to θ and the jump sizes of $\mathcal{A}$ at the observed event times (and also at the observed responses in the case (11)). This maximization can be carried out in many scientific computing packages. For example, the 'Optimization toolbox' of MATLAB (Gilat, 2004) contains an algorithm `fminunc` for unconstrained non-linear optimization. We may choose between large scale and medium scale optimization. The large scale optimization algorithm is a subspace trust region method that is based on the interior reflective Newton algorithm of Coleman and Li (1994, 1996). Each iteration involves approximate solution of a large linear system by using the technique of preconditioned conjugate gradients. The gradient of the function is required. The Hessian matrix is not required and is estimated numerically when it is not supplied. In our implementation, we normally provide the Hessian matrix, so that the algorithm is faster and more reliable. The medium scale optimization is based on the BFGS quasi-Newton algorithm with a mixed quadratic and cubic line search procedure. This algorithm is also available in Press *et al.* (1992). MATLAB also contains an algorithm `fmincon` for constrained non-linear optimization, which is similar to `fminunc`.

The optimization algorithms do not guarantee a global maximum and may be slow for large sample sizes. Our experience, however, shows that these algorithms perform very well for small and moderate sample sizes provided that the initial values are appropriately chosen. We may use the estimates from the Cox model or a parametric model as the initial values. We may also use some other sensible initial values, such as 0 for the regression parameters and Y for $H(Y)$. To gain more confidence in the estimates, one may try different initial values.

It is natural to fit random-effects models through the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977), in which random effects pertain to missing data. The EM algorithm is particularly convenient for the proportional hazards model with random effects because, in the M-step, the estimator of the regression parameter is the root of an estimating function that takes the same form as the partial likelihood score function and the estimator for $\mathcal{A}$ takes the form of the Breslow estimator; see Nielsen *et al.* (1992), Klein (1992) and Andersen *et al.* (1997) for the formulae in the special case of gamma frailty.

For transformation models without random effects, we may use the Laplace transformation to convert the problem into the proportional hazards model with a random effect. Let ξ be a random variable whose density $f(\xi)$ is the inverse Laplace transformation of $\exp\{-G(t)\}$, i.e.

$$\exp\{-G(t)\} = \int_0^\infty \exp(-t\xi)\ f(\xi)\ \mathrm{d}\xi.$$

If

$$P(T > t|\xi) = \exp\left[-\xi \int_0^t \exp\{\beta^{\mathrm{T}} Z(s)\}\ \mathrm{d}\Lambda(s)\right],$$

then

$$P(T > t) = \exp\left(-G\left[\int_0^t \exp\{\beta^{\mathrm{T}} Z(s)\}\ \mathrm{d}\Lambda(s)\right]\right).$$

Thus, we can turn the estimation of the general transformation model into that of the proportional hazards frailty model. This trick also works for general transformation models with random effects, although then there are two sets of random effects in the likelihood; see Appendix A.1 for details.

There is another simple and efficient approach. Using either the forward or the backward recursion that is described in Appendix A.2, we can reduce the task of solving equations for θ and all the jump sizes of Λ to that of solving equations for θ and only one of the jump sizes. This procedure is more efficient and more stable than direct optimization.

4. Asymptotic properties

We consider the general likelihood that is given in equation (12). Denote the true values of θ and $\mathcal{A}$ by θ_0 and $\mathcal{A}_0$ and their NPMLEs by $\hat{\theta}$ and $\hat{\mathcal{A}}$. Under mild regularity conditions, $\hat{\theta}$ is strongly consistent for θ_0 and $\hat{\mathcal{A}}(\cdot)$ uniformly converges to $\mathcal{A}_0(\cdot)$ with probability 1. In addition, the random element $n^{1/2}\{\hat{\theta}-\theta_0, \hat{\mathcal{A}}(\cdot)-\mathcal{A}_0(\cdot)\}$ converges weakly to a zero-mean Gaussian process, and the limiting covariance matrix of $\hat{\theta}$ achieves the semiparametric efficiency bound (Sasieni, 1992; Bickel *et al.*, 1993).

To estimate the variances and covariances of $\hat{\theta}$ and $\hat{\mathcal{A}}(\cdot)$, we treat equation (12) as a parametric likelihood with θ and the jump sizes of $\mathcal{A}$ as the parameters and then invert the observed information matrix for all these parameters. This procedure not only allows us to estimate the covariance matrix of $\hat{\theta}$, but also the covariance function for any functional of $\hat{\theta}$ and $\hat{\mathcal{A}}(\cdot)$. The latter is obtained by the delta method (Andersen *et al.* (1992), section II.8) and is useful in predicting occurrences of events. A limitation of this approach is that it requires inverting a potentially large dimensional matrix and thus may not work well when there are a large number of observed failure times.

When the interest lies primarily in θ, we can use the profile likelihood method (Murphy and van der Vaart, 2000). Let $\mathrm{pl}_n(\theta)$ be the profile log-likelihood function for θ, i.e. $\mathrm{pl}_n(\theta)=\max_{\mathcal{A}}[\log\{L_n(\theta,\mathcal{A})\}]$. Then the (s,t)th element of the inverse covariance matrix of $\hat{\theta}$ can be estimated by

$$-\varepsilon_n^{-2}\{\mathrm{pl}_n(\hat{\theta}+\varepsilon_n e_s+\varepsilon_n e_t)-\mathrm{pl}_n(\hat{\theta}+\varepsilon_n e_s-\varepsilon_n e_t)-\mathrm{pl}_n(\hat{\theta}-\varepsilon_n e_s+\varepsilon_n e_t)+\mathrm{pl}_n(\hat{\theta})\},$$

where ε_n is a constant of order $n^{-1/2}$, and e_s and e_t are the sth and tth canonical vectors respectively. The profile likelihood function can be easily calculated through the algorithms that were described in the previous section. Specifically, $\mathrm{pl}_n(\theta)$ can be calculated via the EM algorithm by holding θ fixed in both the E-step and the M-step. In this way, the calculation is very fast owing to the explicit expression of the estimator of $\mathcal{A}$ in the M-step. In the recursive formulae, the profile likelihood function is a natural product of the algorithm.

The regularity conditions are described in Appendix B. There are three sets of conditions. The first set consists of the compactness of the Euclidean parameter space, the boundedness of covariates, the non-emptyness of risk sets and the boundedness of the number of events (i.e. conditions D1–D4 in Appendix B); these are standard assumptions for any survival analysis and are essentially the regularity conditions of Andersen and Gill (1982). The second set of conditions pertains to the transformation function and random effects (i.e. conditions D5 and D6); these conditions hold for all commonly used transformation functions and random-effects distributions. The final set of conditions pertains to the identifiability of parameters (i.e. conditions D7 and D8); these conditions hold for the models and data structures that are considered in this paper provided that the covariates are linearly independent and the distribution of the random effects has a unique parameterization. In short, the regularity conditions hold in all practically important situations.

5. Examples

5.1. Gastrointestinal tumour study

As mentioned previously, Stablein and Koutrouvelis (1985) presented survival data from a clinical trial on locally unresectable gastric cancer. Half of the total 90 patients were assigned to chemotherapy, and the other half to combined chemotherapy and radiotherapy. There were two censored observations in the first treatment arm and six in the second. Under the two-sample proportional hazards model, the log-hazard ratio is estimated at 0.106 with a standard error estimate of 0.223, yielding a p-value of 0.64. This analysis is meaningless in view of the crossing survival curves that are shown in Fig. 1.

We fit models (3) and (4) with $G(x)=x$ and $Z\equiv\tilde{Z}$ indicating chemotherapy *versus* combined therapy by the values 1 *versus* 0. We use the backward recursive formula of Appendix A.2 to calculate the NPMLEs. Under model (3), β and γ are estimated at 0.317 and −0.530 with standard error estimates of 0.190 and 0.093. Under model (4), the estimates of β and γ become 3.028 and −1.317 with standard error estimates of 0.262 and 0.032. As evident in Fig. 1, model (4) fits the data better than model (3) and accurately reflects the observed pattern of crossing survival curves.

5.2. Colon cancer study

In the colon cancer study that was mentioned in Section 1, 315, 310 and 304 patients with stage C disease received observation, levamisole alone and levamisole combined with 5-fluorouracil (group Lev+5-FU) respectively. By the end of the study, 155 patients in the observation group, 144 in the levamisole alone group and 103 in the Lev+5-FU group had recurrences of cancer, and there were 114, 109 and 78 deaths in the observation, levamisole alone and Lev+5-FU groups respectively. Lin (1994) fitted separate proportional hazards models to recurrence of cancer and death. That analysis ignored the informative censoring on cancer recurrence and did not explore the joint distribution of the two end points.

Following Lin (1994), we focus on the comparison between the observation and Lev+5-FU groups. We treat recurrence of cancer as the first type of failure and death as the second, and we consider four covariates:

$$Z_{1i}=\begin{cases}0 & \text{if the } i\text{th patient was on observation,}\\ 1 & \text{if the } i\text{th patient was on Lev+5-FU;}\end{cases}$$

$$Z_{2i}=\begin{cases}0 & \text{if the surgery for the } i\text{th patient took place 20 or fewer days before randomization,}\\ 1 & \text{if the surgery for the } i\text{th patient took place more than 20 days before randomization;}\end{cases}$$

$$Z_{3i}=\begin{cases}0 & \text{if the depth of invasion for the } i\text{th patient was submucosa or muscular layer,}\\ 1 & \text{if the depth of invasion for the } i\text{th patient was serosa;}\end{cases}$$

$$Z_{4i}=\begin{cases}0 & \text{if the number of nodes involved in the } i\text{th patient was 1–4,}\\ 1 & \text{if the number of nodes involved in the } i\text{th patient was more than 4.}\end{cases}$$

We fit the class of models in equation (7) with a normal random-effect and the Box–Cox transformations $\{(1+x)^{\rho}-1\}/\rho$ and logarithmic transformations $r^{-1}\log(1+rx)$ through the EM algorithm. The log-likelihood functions under these transformations are shown in Fig. 2. The combination of $G_1(x)=2\{(1+x)^{1/2}-1\}$ and $G_2(x)=\log(1+1.45x)/1.45$ maximizes the

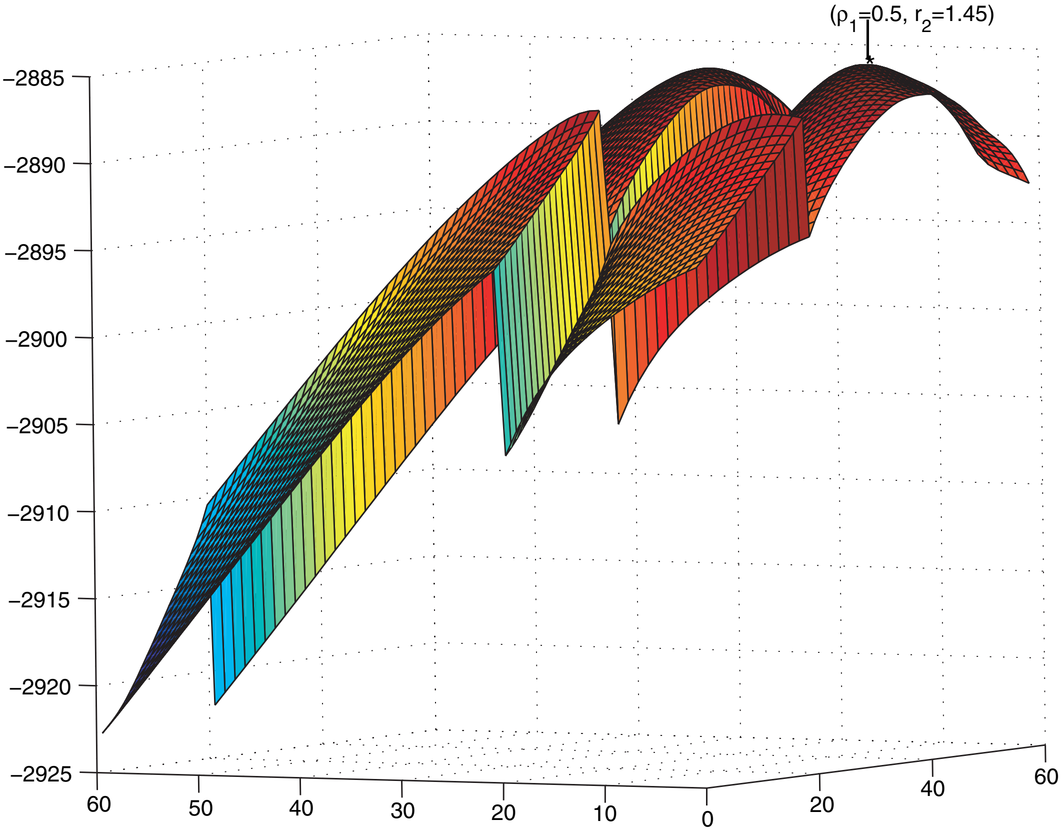

Fig. 2. Log-likelihood functions for pairs of transformations in the colon cancer data: indices below 20 pertain to the Box–Cox transformations with ρ ranging from 1 to 0, whereas indices above 20 pertain to the logarithmic transformations with r ranging from 0 to 2

likelihood function. By the Akaike (1985) information criterion, we select this bivariate model. Table 1 presents the results under the model selected and the proportional hazards and proportional odds models. All three models show that the Lev+5-FU treatment is effective in preventing recurrence of cancer and death. The interpretation of treatment effects and the prediction of events depend on which model is used.

We can predict an individual's future events on the basis of his or her event history. The survival probability at time t for a patient with covariate values z and with cancer recurrence at t_0 is

$$\left(\int_b \exp[-G_2\{\Lambda_2(t)\exp(\beta_2^{\mathrm{T}}z+b)\}]\,G_1'\{\Lambda_1(t_0)\exp(\beta_1^{\mathrm{T}}z+b)\}\exp[-G_1\{\Lambda_1(t_0)\right.$$
$$\left.\times\exp(\beta_1^{\mathrm{T}}z+b)\}]\,\mathrm{d}\Phi(b/\sigma_b)\right)\left(\int_b \exp[-G_2\{\Lambda_2(t_0)\exp(\beta_2^{\mathrm{T}}z+b)\}]\,G_1'\{\Lambda_1(t_0)\exp(\beta_1^{\mathrm{T}}z+b)\}\right.$$
$$\left.\times\exp[-G_1\{\Lambda_1(t_0)\exp(\beta_1^{\mathrm{T}}z+b)\}]\,\mathrm{d}\Phi(b/\sigma_b)\right)^{-1}, \qquad t\geqslant t_0,$$

where Φ is the standard normal distribution function. We estimate this probability by replacing all the unknown parameters with their sample estimators and estimate the standard error by the delta method. An example of this kind of prediction is given in Fig. 3.

Table 1. Estimates of regression parameters and variance component under random-effects transformation models for the colon cancer study†

	Estimates for the following models:		
	Proportional hazards	*Proportional odds*	*Selected*
Treatment			
Cancer	−1.480 (0.236)	−1.998 (0.352)	−2.265 (0.357)
Death	−0.721 (0.282)	−0.922 (0.379)	−1.186 (0.422)
Surgery			
Cancer	−0.689 (0.219)	−0.786 (0.335)	−0.994 (0.297)
Death	−0.643 (0.258)	−0.837 (0.369)	−1.070 (0.366)
Depth			
Cancer	2.243 (0.412)	3.012 (0.566)	3.306 (0.497)
Death	1.937 (0.430)	2.735 (0.630)	3.033 (0.602)
Node			
Cancer	2.891 (0.236)	4.071 (0.357)	4.309 (0.341)
Death	3.095 (0.269)	4.376 (0.384)	4.742 (0.389)
σ_b^2	11.62 (1.22)	24.35 (2.46)	28.61 (3.06)
Log-likelihood	−2895.1	−2895.0	−2885.7

†Standard error estimates are shown in parentheses.

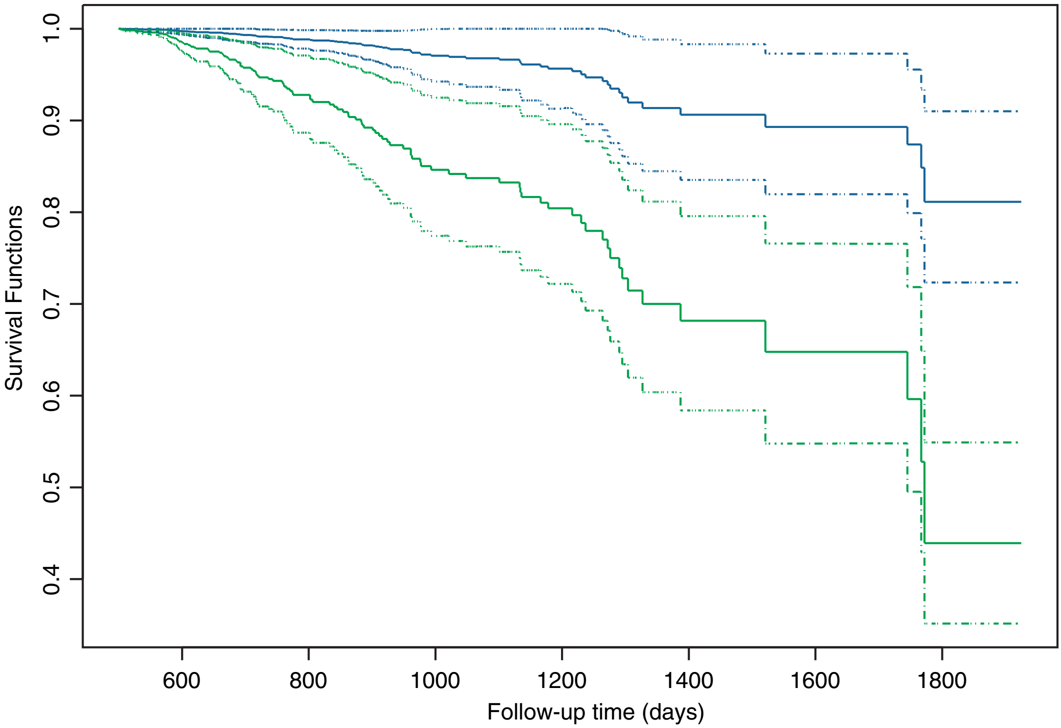

Fig. 3. Estimated survival probabilities of the colon cancer patients with recurrences of cancer at 500 days under the model selected (the blue and green curves pertain to $z=(1,1,0,0)$ and $z=(0,0,1,1)$ respectively): ———, point estimates; - · - · - ·, pointwise 95% confidence limits

To test the global null hypothesis of no treatment effect on recurrence of cancer and death, we may impose the condition of a common treatment effect while allowing separate effects for the other covariates. The estimates of the common treatment effects are -1.295, -1.523 and -1.843, with standard error estimates of 0.256, 0.333 and 0.318 under the proportional hazards, proportional odds and selected models. Thus, we would conclude that the Lev+5-FU treatment is highly efficacious.

5.3. Human immunodeficiency virus study

A clinical trial was conducted to evaluate the benefit of switching from zidovudine to didanosine (ddI) for HIV patients who have tolerated zidovudine for at least 16 weeks (Lin and Ying, 2003). A total of 304 patients were randomly chosen to continue the zidovudine therapy whereas 298 patients were assigned to ddI. The investigators were interested in comparing the CD4 cell counts between the two groups at weeks 8, 16 and 24. A total of 174 zidovudine patients and 147 ddI patients dropped out of the study owing to patient's request, physician's decision, toxicities, death and other reasons.

To adjust for informative drop-out in the analysis of CD4 cell counts, we use a special case of equation (10):

$$\tilde{H}(Y_{ij}) = \alpha_1 X_i + \alpha_2 t_{ij} + b_i + \varepsilon_{ij}, \qquad (13)$$

where X_i is the indicator for ddI, t_{ij} is 8, 16 and 24 weeks, b_i is zero-mean normal with variance σ_b^2 and ε_{ij} is standard normal. Table 2 summarizes the results of this analysis, along with the results based on the log- and square-root transformations. These results indicate that ddI slowed down the decline of CD4 cell counts over time. The analysis that is based on the estimated transformation provides stronger evidence for the ddI effect than those based on the parametric transformations. Model (13) includes the random intercept; additional analysis reveals that the random slope is not significant.

Fig. 4 suggests that neither the log- nor the square-root transformation provides a satisfactory approximation to the true transformation. The histograms of the residuals (which are not shown here) reveal that the residual distribution is normal looking under the estimated transformation, is right skewed under the square-root transformation and left skewed under the log-transfor-

Table 2. Joint analysis of CD4 cell counts and drop-out time for the HIV study†

Parameter	*Results for the following transformation functions:*					
	Estimated		*Logarithmic*		*Square root*	
	Est	*SE*	*Est*	*SE*	*Est*	*SE*
α_1	0.674	0.222	0.506	0.215	0.613	0.261
α_2	−0.043	0.005	−0.041	0.005	−0.041	0.004
β	−0.338	0.114	−0.316	0.116	−0.328	0.118
σ_b^2	7.837	0.685	7.421	0.575	8.994	0.772
ψ	−0.158	0.023	−0.132	0.021	−0.154	0.023

†The parameters α_1 and α_2 represent the effects of ddI and time on CD4 cell counts, and β pertains to the effect of ddI on the time to drop-out. The estimates of α, σ_b^2 and ψ under the log- and square-root transformations are standardized to have unit residual variance. Est and SE denote the parameter estimate and standard error estimate.

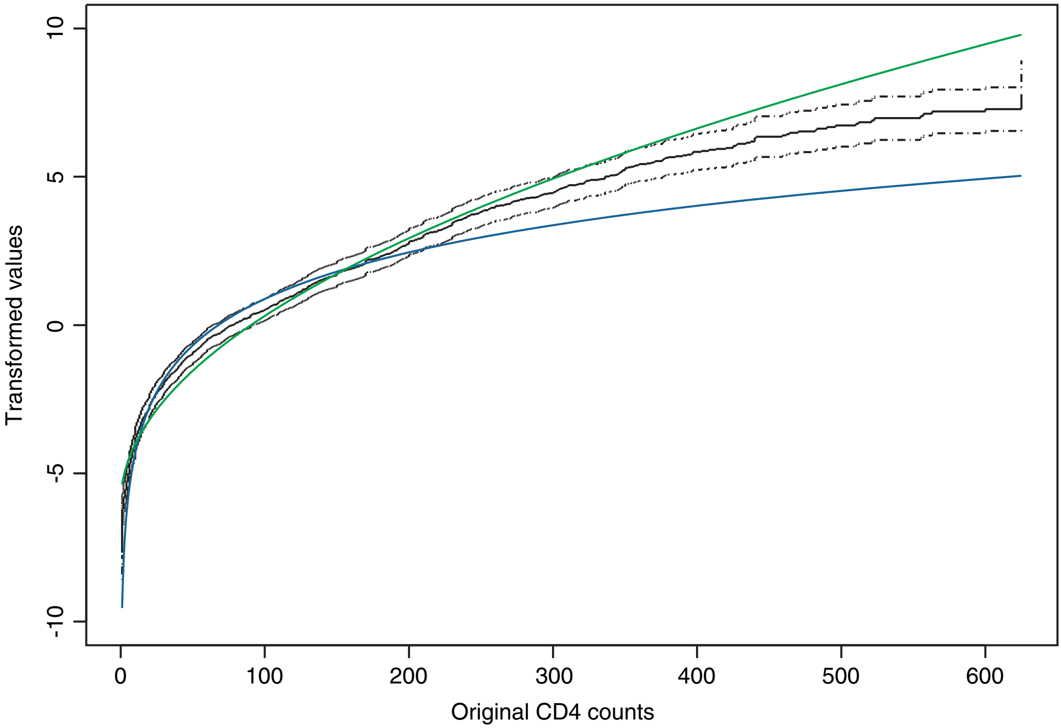

Fig. 4. Transformation functions for the HIV study (the blue and green curves pertain respectively to the log- and square-root transformation functions subject to affine transformations): ——, estimated transformation function; - · - · - ·, corresponding pointwise 95% confidence limits

mation. In addition, the *qq*-norm plots of the residuals (which are not shown) indicate that the estimated transformation is much more effective in handling the extreme observations than the log- and square-root transformations.

Without adjustment of informative drop-out, the estimates of α_1 and α_2 under model (13) shrink drastically to 0.189 and -0.011. The same model is used for CD4 cell counts in the two analyses, but the estimators are severely biased when informative drop-out is not accounted for.

6. Simulation studies

We conducted extensive simulation studies to assess the performance of the inferential and numerical procedures proposed. The first set of studies mimicked the colon cancer study. We generated two types of failures with cumulative hazard functions $G_k\{\exp(\beta_{1k}Z_{1i}+\beta_{2k}Z_{2i}+b_i)\times \Lambda_k(t)\}$ $(k=1,2;\ i=1,\ldots,n)$, where Z_{1i} and Z_{2i} are independent Bernoulli and uniform [0,1] variables, b_i is standard normal, $\beta_{11}=\beta_{12}=-\beta_{21}=-\beta_{22}=1$, $\Lambda_1(t)=0.3t$, $\Lambda_2(t)=0.15t^2$ and $G_1(x)=G_2(x)$ equals x or $\log(1+x)$. We created censoring times from the uniform [0, 5] distribution and set $\tau=4$, producing approximately 51.3% and 48.5% censoring for $k=1$ and $k=2$ under $G_1(x)=G_2(x)=x$, and 59.9% and 57.3% under $G_1(x)=G_2(x)=\log(1+x)$. We used the EM algorithm that is described in Appendix A.1 to calculate the NPMLEs.

Table 3 summarizes the results for β_{11}, β_{21}, $\Lambda_1(t)$ and σ_b^2, where σ_b^2 is the variance of the random effect. The results for β_{12}, β_{22} and $\Lambda_2(t)$ are similar and have been omitted. The estimators of β_k appear to be virtually unbiased. There are some biases for the estimator of σ_b^2 and for the estimator of $\Lambda_k(t)$ near the right-hand tail, although the biases decrease rapidly with sample

Table 3. Simulation results for bivariate failure time data†

n	*Parameter*	*Results for $G_1(x)=G_2(x)=x$*				*Results for $G_1(x)=G_2(x)$ $=log(1+x)$*			
		Bias	*SE*	*SEE*	*CP*	*Bias*	*SE*	*SEE*	*CP*
100	β_{11}	−0.014	0.406	0.392	0.942	−0.013	0.516	0.496	0.947
	β_{21}	0.025	0.671	0.664	0.955	0.030	0.849	0.847	0.957
	σ_b^2	−0.089	0.489	0.482	0.965	−0.125	0.604	0.717	0.963
	$\Lambda_1(\tau/4)$	0.030	0.156	0.145	0.945	0.043	0.210	0.190	0.952
	$\Lambda_1(3\tau/4)$	0.073	0.474	0.429	0.952	0.116	0.655	0.571	0.955
200	β_{11}	0.000	0.286	0.277	0.949	0.003	0.395	0.350	0.950
	β_{21}	0.007	0.474	0.468	0.948	0.016	0.599	0.596	0.955
	σ_b^2	−0.037	0.353	0.346	0.961	−0.054	0.468	0.509	0.957
	$\Lambda_1(\tau/4)$	0.014	0.104	0.099	0.944	0.018	0.130	0.126	0.953
	$\Lambda_1(3\tau/4)$	0.032	0.305	0.291	0.948	0.044	0.393	0.375	0.952
400	β_{11}	−0.000	0.207	0.196	0.943	−0.002	0.258	0.247	0.940
	β_{21}	0.009	0.329	0.331	0.952	0.014	0.417	0.420	0.950
	σ_b^2	−0.011	0.251	0.247	0.961	−0.024	0.335	0.362	0.959
	$\Lambda_1(\tau/4)$	0.005	0.070	0.069	0.948	0.008	0.088	0.087	0.954
	$\Lambda_1(3\tau/4)$	0.014	0.210	0.202	0.947	0.020	0.267	0.259	0.950

†Bias and SE are the bias and standard error of the parameter estimator, SEE is the mean of the standard error estimator and CP is the coverage probability of the 95% confidence interval. The confidence intervals for $\Lambda(t)$ are based on the log-transformation, and the confidence interval for σ_b^2 is based on the Satterthwaites (1946) approximation. Each entry is based on 5000 replicates.

size. The variance estimators are fairly accurate, and the confidence intervals have reasonable coverage probabilities.

In the second set of studies, we generated recurrent event times from the counting process with cumulative intensity $G\{\Lambda(t)\exp(\beta_1 Z_1+\beta_2 Z_2+b)\}$, where Z_1 is Bernoulli with 0.5 success probability, Z_2 is normal with mean Z_1 and variance 1, b is normal with mean 0 and variance σ_b^2, $\Lambda(t)=\lambda\log(1+t)$ and $G(x)=\{(1+x)^\rho-1\}/\rho$ or $G(x)=\log(1+rx)/r$. We generated censoring times from the uniform [2, 6] distribution and set τ to 4. We considered various choices of β_1, β_2, ρ, r, λ and σ_b^2. We used a combination of the EM algorithm and the backward recursive formula to calculate the NPMLEs. The results are very similar to those of Table 3 and thus have been omitted.

The third set of studies mimicked the HIV study. We generated repeated measures from model (13), in which X_i is Bernoulli with 0.5 success probability and $t_{ij}=j\tau/5$ $(j=1,\ldots,4)$. We set $\tilde{H}(y)=\log(y)$ or

$$\tilde{H}(y)=\log\left\{\frac{(1+y)^2-1}{2}\right\},$$

and let the transformation function be unspecified in the analysis. We generated survival times from the proportional hazards model with conditional hazard function $0.3t\exp(\beta X_i+\psi b_i)$, and censoring times from the uniform [0, 5] distribution with $\tau=4$. The censoring rate was approximately 53%, and the average number of repeated measures was about 1.58 per subject. We used the optimization algorithm `fminunc` in MATLAB to obtain the NPMLEs. We penalized the objective function for negative estimates of variance and jump sizes by setting its value to -10^6. The results are similar to those of the first two sets of studies.

Table 4. Simulation results for joint modelling of repeated measures and survival time†

n	*Parameter*	*Results for $H(y)=\log(y)$*				*Results for $H(y)=\log[\{(1+y)^2-1\}/2]$*			
		Bias	*SE*	*SEE*	*CP*	*Bias*	*SE*	*SEE*	*CP*
100	α_1	−0.020	0.253	0.248	0.941	−0.017	0.250	0.249	0.943
	α_2	−0.011	0.207	0.203	0.946	−0.011	0.208	0.205	0.947
	β	−0.041	0.415	0.416	0.960	−0.047	0.415	0.418	0.959
	σ_1^2	−0.063	0.403	0.415	0.963	−0.054	0.400	0.418	0.965
	σ_2^2	−0.053	0.568	0.570	0.935	−0.036	0.553	0.580	0.949
	ψ_1	0.082	0.453	0.550	0.956	0.084	0.463	0.553	0.969
	ψ_2	0.022	0.514	0.602	0.967	0.013	0.501	0.612	0.983
	$\tilde{\Lambda}(1)$	0.015	0.201	0.196	0.947	0.016	0.308	0.302	0.948
	$\tilde{\Lambda}(3)$	0.027	0.730	0.692	0.940	0.082	2.488	2.394	0.939
	$\Lambda(3\tau/4)$	0.006	0.180	0.172	0.954	0.008	0.180	0.173	0.954
200	α_1	−0.013	0.177	0.176	0.948	−0.014	0.177	0.176	0.948
	α_2	−0.007	0.145	0.145	0.949	−0.007	0.145	0.145	0.949
	β	−0.028	0.278	0.283	0.960	−0.028	0.279	0.283	0.960
	σ_1^2	−0.041	0.297	0.301	0.967	−0.042	0.297	0.301	0.966
	σ_2^2	−0.047	0.411	0.411	0.958	−0.048	0.412	0.411	0.957
	ψ_1	0.053	0.322	0.351	0.969	0.053	0.322	0.341	0.968
	ψ_2	0.014	0.351	0.366	0.979	0.014	0.351	0.366	0.979
	$\tilde{\Lambda}(1)$	0.009	0.140	0.138	0.950	0.008	0.215	0.212	0.947
	$\tilde{\Lambda}(3)$	0.012	0.493	0.485	0.950	0.022	1.696	1.651	0.943
	$\Lambda(3\tau/4)$	0.002	0.122	0.118	0.950	0.002	0.122	0.118	0.950

†Bias and SE are the bias and standard error of the parameter estimator, SEE is the mean of the standard error estimator and CP is the coverage probability of the 95% confidence interval. The confidence intervals for $\tilde{\Lambda}(t)$ are based on the log-transformation, and the confidence intervals for σ_1^2 and σ_2^2 are based on the Satterthwaites (1946) approximation. Each entry is based on 5000 replicates.

The fourth set of studies is the same as the third except that the scalar random effect b_i on the right-hand side of equation (13) is replaced by $b_{1i}+b_{2i}t_{ij}$. The random effects b_{1i} and b_{2i} enter the survival time model with coefficients ψ_1 and ψ_2 respectively. We generated $(b_{1i},b_{2i})^{\mathrm{T}}$ from the zero-mean normal distribution with variances σ_1^2 and σ_2^2 and covariance σ_{12}. Table 4 reports the results for $\alpha_1=1$, $\alpha_2=-\beta=0.5$, $\psi_1=1$, $\psi_2=0.5$, $\sigma_1^2=\sigma_2^2=1$ and $\sigma_{12}=-0.4$. We again conclude that the asymptotic approximations are sufficiently accurate for practical use.

In the first three sets of studies, which involve scalar random effects, it took about 5 s on an IBM BladeCenter HS20 machine to complete one simulation with $n=200$. In the fourth set of studies, which involves two random effects, it took about 7 min and 35 min to complete one simulation with $n=100$ and $n=200$ respectively. In the first three sets of studies, the algorithms failed to converge on very rare occasions with $n=100$ and always converged with $n=200$ and $n=400$. In the fourth set of studies, the algorithm failed in about 0.4% occasions with $n=100$ and 0.2% of the time with $n=200$.

We conducted additional studies to compare the methods proposed with the existing methods. For the class of models in equation (1), the best existing estimators are those of Chen *et al.* (2002). We generated survival times with cumulative hazard rate

$$\log[1+r\{\Lambda(t)\exp(\beta_1 Z_1+\beta_2 Z_2)\}]/r,$$

where Z_1 is Bernoulli with 0.5 success probability, Z_2 is normal with mean Z_1 and unit variance, $\Lambda(t)=3t$, $\beta_1=-1$ and $\beta_2=0.2$. We simulated exponential censoring times with a hazard rate that was chosen to yield a desired level of censoring under $\tau = 6$. Our algorithm always converged, whereas the program that was kindly provided by Z. Jin failed to converge in about 2% of the simulated data sets. For $n = 100$ and 25% censoring, the efficiencies of the estimators of Chen *et al.* (2002) relative to the NPMLEs are approximately 0.92, 0.83 and 0.69 under $r = 0.5$, 1, 2 respectively, for both β_1 and β_2. We also compared the estimators proposed with those of Cai *et al.* (2002) for clustered failure time data and found that the former are much faster to compute and considerably more efficient than the latter; see Zeng *et al.* (2005) for the specific results under the proportional odds model with normal random effects.

7. Discussion

The present work contributes to three aspects of semiparametric regression models with censored data. First, we present several important extensions of the existing models. Secondly, we develop a general asymptotic theory for the NPMLEs of such models. Thirdly, we provide simple and efficient numerical methods to implement the corresponding inference procedures. We hope that our work will facilitate further development and applications of semiparametric models.

In the transformation models, the function G is regarded as fixed. One may specify a parametric family of functions and then estimate the relevant parameters. This is in a sense what we did in Section 5.2, but we did not account for the extra variation that is due to the estimation of those parameters. It is theoretically possible, although computationally demanding, to account for the extra variation. Whether this kind of variation should be accounted for is debatable (Box and Cox, 1982). Leaving G non-parametric is a challenging topic that is currently being pursued by statisticians and econometricians.

As argued in Sections 1, 2.3 and 5.3, it is desirable to use the semiparametric linear mixed model that is given in equation (10) so that parametric transformation can be avoided. It is surprising that this model has not been proposed earlier. Our simulation results (which are not shown here) reveal that the NPMLEs of the regression parameters and variance components are nearly as efficient as if the true transformation were known. Thus, we recommend that semiparametric linear regression be adopted for both single and repeated measures of continuous response, whether or not there is informative drop-out.

In the joint modelling, repeated measures are assumed to be independent conditional on the random effects. One may incorporate a within-subject autocorrelation structure in the model, as suggested by Henderson *et al.* (2000) and Xu and Zeger (2001). One may also use joint models for repeated measures of multiple outcomes. The likelihood functions under such extensions can be constructed. The likelihood approach can handle random intermittent missing values, but not non-ignorable missingness.

The asymptotic theory that is described in Appendix B is very general and can be applied to a large spectrum of semiparametric models with censored data. In the existing literature, the asymptotic theory for the NPMLE has been proved case by case only. This kind of proof involves very advanced mathematical arguments. The general theorems that are given in Appendix B enable one to establish the desired asymptotic results for a specific problem by checking a few regularity conditions, which is much easier than proving the results from scratch.

There are some gaps in the theory. First, we have been unable to prove the asymptotics of the NPMLEs for linear transformation models completely when the observations on the response variable are unbounded. This means that the NPMLE for model (10) does not yet have rigorous theoretical justifications, although the desired asymptotic properties are strongly supported by our simulation results. Secondly, there is no proof in the literature for the asymptotic distribution of the likelihood ratio statistic under a semiparametric model when the parameter of interest lies on the boundary of the parameter space. This is a serious deficiency since we might want to test the hypothesis of zero variance in random-effects models. In many parametric cases, the limiting distributions of likelihood ratio statistics are mixtures of χ^2-distributions (Self and Liang, 1987). We expect those results to hold for the kind of semiparametric model that is considered in this paper. This conjecture is well supported by our simulation results (e.g. Diao and Lin (2005)), although it remains to be proved.

The counting process martingale theory, which has been the workhorse behind the theoretical development of survival analysis over the last quarter of a century, plays no role in establishing the asymptotic theory for the kind of problem that is considered in this paper, not even for univariate survival data. We have relied heavily on modern empirical process theory, which we believe will be the primary mathematical tool in survival analysis and semiparametric inference more broadly for the foreseeable future.

The EM algorithms that are described in Appendix A.1 are similar to the QEM algorithm of Tsodikov (2003), but the latter is confined to univariate failure time data. Although we have very good experience with them, the convergence rates of such semiparametric EM algorithms have not been investigated in the literature. It is unclear whether the recursive formulae that are given in Appendix A.2 are applicable to time varying covariates. Whether the Laplace transformation idea that is described in Section 3 can be extended to recurrent events is also an open question. Thus, the extent to which the NPMLEs will be generally adopted depends on further advances in numerical algorithms.

It is desirable to choose the 'best' model among all possible ones. We used the Akaike information criterion to select the transformations in Section 5.2. A related method is the Bayesian information criterion (Schwarz, 1978). An alternative approach is likelihood-based cross-validation. Another strategy is to formalize the prediction error criterion that was used in Section 5.1. Further research is warranted.

We have demonstrated through three types of problem that the NPMLE is a very general and powerful approach to the analysis of semiparametric regression models with censored data. This approach can be used to study many other problems. We list below some potential areas of research.

7.1. Cure models

In some applications, a proportion of the subjects may be considered cured in that they will not experience the event of interest even after extended follow-up (Farewell, 1982). Peng and Dear (2000) and Sy and Taylor (2000) described EM algorithms for computing the NPMLEs for a mixture cure model that postulates a proportional hazards model for the susceptible individuals, but they did not study their theoretical properties. It is desirable to extend this model by replacing the proportional hazards model with the class of transformation models that is given in equations (2) or (3), to allow non-proportional hazards models and recurrent events. The asymptotic properties are expected to follow from the general theorems of Appendix B, although the conditions need to be verified.

7.2. Joint models for recurrent and terminal events

In many instances, the observation of recurrent events is ended by a terminal event, such as death or drop-out. Shared random-effects models which are similar to those described in Section 2.3 have been proposed to formulate the joint distribution of recurrent and terminal events (e.g. Wang *et al.* (2001), Liu *et al.* (2004) and Huang and Wang (2004)). In particular, Liu *et al.* (2004) incorporated a common gamma frailty into the proportional intensity model for the recurrent events and the proportional hazards model for the terminal event. They developed a Monte Carlo EM algorithm to obtain the NPMLEs but provided no theoretical justifications. One may extend the joint model of Liu *et al.* (2004) by replacing the proportional hazards or intensity model with the general random-effects transformation models and try to establish the asymptotic properties of the NPMLEs by appealing to the general theorems of Appendix B.

7.3. Missing covariates

Robins *et al.* (1994) and Nan *et al.* (2004) obtained the information bounds with missing data. Chen and Little (1999) and Chen (2002) studied the NPMLEs for the proportional hazards model with missing covariates, whereas Scheike and Juul (2004) and Scheike and Martinussen (2004) considered the specific situations in which covariates are missing because of case–cohort or nested case–control sampling (Kalbfleisch and Prentice (2002), page 339). To make the NPMLEs tractable, one normally assumes data missing at random and imposes certain restrictions on the covariate distribution. How general the covariate distribution can be is an open question.

7.4. Genetic studies

Models (7) and (10) can be extended to genetic linkage and association studies on potentially censored non-normal quantitative traits, whereas models (2) and (7) can be adapted to haplotype-based association studies (e.g. Diao and Lin (2005) and Lin and Zeng (2006)); inference on haplotype–disease association, which is a hot topic in genetics, is essentially a missing or mismeasured covariate problem. The analysis of genetic data by the NPMLE is largely uncharted.

There are alternative approaches to the NPMLE. Martingale-based estimating equations were used by Chen *et al.* (2002) for linear transformation models and by Lu and Ying (2004) for cure models. This approach can also be applied to the general transformation models that are given in equations (2)–(4). The inverse probability of censoring weighting (Robins and Rotnitzky, 1992) approach was used by Cheng *et al.* (1995) and Cai *et al.* (2002) for linear transformation models, and by Kalbfleisch and Lawless (1988), Borgan *et al.* (2000) and Kulich and Lin (2004) for case–cohort studies. These estimators are not asymptotically efficient. The estimating equations are usually solved by Newton–Raphson algorithms, which may not converge. The moment-based estimators are expected to be more robust than the NPMLEs against model misspecification. It would be worthwhile to assess the robustness *versus* efficiency of the two approaches through simulation studies.

Marginal models (Wei *et al.* (1989) and Kalbfleisch and Prentice (2002), pages 305–306) are used almost exclusively in the analysis of multivariate failure time data, mainly because of their robustness and available commercial software. Because in general marginal and random-effects models cannot hold simultaneously, there is a debate about which approach is more meaningful. Random-effects models have important advantages. First, they enable us to predict future events on the basis of an individual's event history, as shown in Fig. 3, or to predict a person's survival outcome given the survival times of other members of the same cluster. Secondly, they allow efficient parameter estimation. Thirdly, the dependence structures are of scientific interest in many applications, especially in genetics.

Our work does not cover the accelerated failure time model, which takes the form of equation (1) but with known H and unknown distribution of ε (Kalbfleisch and Prentice (2002), pages

218–219). Rank and least squares estimators for this model have been studied extensively over the last three decades; see Kalbfleisch and Prentice (2002), chapter 7. These estimators are not asymptotically efficient. In addition, it is difficult to calculate them or to estimate their variances, although progress has been made on this front (Jin *et al.*, 2003, 2006). We are pursuing a variant of the NPMLE for the accelerated failure time model with potentially time varying covariates, which maximizes a kernel-smoothed profile likelihood function. The estimator is consistent, asymptotically normal and asymptotically efficient with an easily estimated variance, and it works well in real and simulated data.

We have focused on right-censored data. Interval censoring arises when the failure time is only known to fall in some interval. It is much more challenging to apply the NPMLE to interval-censored data than to right-censored data. So far asymptotic theory is only available for proportional hazards models with current status data (Huang, 1996), which arise when the failure time is only known to be less than or greater than a single monitoring time. Wellner and Zhang (2005) studied proportional mean models for panel counts data with general interval censoring. We expect considerable theoretical and numerical innovation in this area in the coming years.

We have taken a frequentist approach. Ibrahim *et al.* (2001) provided an excellent description of Bayesian methods for semiparametric models with censored data. There are many recent references. It would be valuable to develop the Bayesian counterparts of the methods that were presented in this paper.

Much of the theoretical and methodological development in survival analysis over the last three decades has been centred on the proportional hazards model. Because everything that has been written about that model is also relevant to transformation models, opportunities for research abound. Besides the problems that have already been mentioned earlier, it would be worthwhile to develop methods for variable selection, model checking and robust inference (under misspecified models) and to explore the use of these models in the areas of diagnostic medicine, sequential clinical trials, causal inference, multistate processes, spatially correlated failure time data, and so on.

Acknowledgements

We thank David Cox, Jack Cuzick, Vern Farewell, David Oakes, Peter Sasieni, Richard Smith and Bruce Turnbull, the referees and the Research Section Committee for helpful comments and constructive suggestions. This work was supported by the National Institutes of Health.

Appendix A: Numerical methods

A.1. EM algorithms

We describe an EM algorithm for maximizing the likelihood function that is given in expression (8). Similar algorithms can be used for the other likelihood functions. For simplicity of description, we focus on multiple-events data. The data consist of $(Y_{ik}, \Delta_{ik}, Z_{ik})$ $(i=1,\ldots,n;\ k=1,\ldots,K)$, where Y_{ik} is the observation time for the kth event on the ith subject, Δ_{ik} indicates, by the values 1 *versus* 0, whether Y_{ik} is an uncensored or censored observation and Z_{ik} is the corresponding covariate vector. We wish to maximize the objective function

$$\prod_{i=1}^{n} \int_{b} \prod_{k=1}^{K} \left(\Lambda_k\{Y_{ik}\} \exp\{\beta^{\mathrm{T}} Z_{ik}(Y_{ik}) + b^{\mathrm{T}} \tilde{Z}_{ik}(Y_{ik})\} G'_k \left[\int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}} Z_{ik}(s) + b^{\mathrm{T}} \tilde{Z}_{ik}(s)\} \, \mathrm{d}\Lambda_k(s) \right] \right)^{\Delta_{ik}}$$

$$\times \exp\left(-G_k \left[\int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}} Z_{ik}(t) + b^{\mathrm{T}} \tilde{Z}_{ik}(t)\} \, \mathrm{d}\Lambda_k(t) \right] \right) f(b;\gamma) \, \mathrm{d}b.$$

For all commonly used transformations, including the classes of Box–Cox transformations and logarithmic transformations, $\exp\{-G_k(x)\}$ is the Laplace transformation of some function $\phi_k(x)$ such that

$$\exp\{-G_k(x)\} = \int_0^\infty \exp(-xt)\ \phi_k(t)\ \mathrm{d}t.$$

Clearly, $\int_0^\infty \phi_k(t)\ \mathrm{d}t = 1$. We introduce a new frailty ξ_{ik} with density function ϕ_k. Since

$$G'_k(x) \exp\{-G_k(x)\} = \int_0^\infty t \exp(-xt)\ \phi_k(t)\ \mathrm{d}t,$$

the objective function can be written as

$$\prod_{i=1}^n \int_b \prod_{k=1}^K \int_{\xi_{ik}} [\xi_{ik}\ \Lambda_k\{Y_{ik}\} \exp\{\beta^{\mathrm{T}}\ Z_{ik}(Y_{ik}) + b^{\mathrm{T}}\ \tilde{Z}_{ik}(Y_{ik})\}]^{\Delta_{ik}} \exp\left[-\xi_{ik} \int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}}\ Z_{ik}(t) + b^{\mathrm{T}}\ \tilde{Z}_{ik}(t)\}\ \mathrm{d}\Lambda_k(t)\right] \times \phi_k(\xi_{ik})\ f(b;\gamma)\ \mathrm{d}\xi_{ik}\ \mathrm{d}b.$$

This expression is the likelihood function under the proportional hazards frailty model with conditional hazard function $\xi_{ik}\ \lambda_k(t) \exp\{\beta^{\mathrm{T}}\ Z_{ik}(t) + b_i^{\mathrm{T}}\ \tilde{Z}_{ik}(t)\}$. Thus, treating the b_i and ξ_{ik} as missing data, we propose the following EM algorithm to calculate the NPMLEs.

In the M-step, we solve the complete-data score equation conditional on the observed data. Specifically, we solve the following equation for β:

$$\sum_{i=1}^n \sum_{k=1}^K \Delta_{ik} \left(Z_{ik}(Y_{ik}) - \frac{\sum_{j=1}^n I(Y_{jk} \geqslant Y_{ik})\ Z_{jk}(Y_{ik}) \hat{E}[\xi_{jk} \exp\{\beta^{\mathrm{T}}\ Z_{jk}(Y_{ik}) + b_j^{\mathrm{T}}\ \tilde{Z}_{jk}(Y_{ik})\}]}{\sum_{j=1}^n I(Y_{jk} \geqslant Y_{ik}) \hat{E}[\xi_{jk} \exp\{\beta^{\mathrm{T}}\ Z_{jk}(Y_{ik}) + b_j^{\mathrm{T}}\ \tilde{Z}_{jk}(Y_{ik})\}]} \right) = 0,$$

where $\hat{E}[\cdot]$ is the conditional expectation given the observed data and the current parameter estimates. In addition, we estimate Λ_k as a step function with the following jump size at Y_{ik}:

$$\Delta_{ik} \Big/ \sum_{j=1}^n I(Y_{jk} \geqslant Y_{ik}) \hat{E}[\xi_{jk} \exp\{\beta^{\mathrm{T}}\ Z_{jk}(Y_{ik}) + b_j^{\mathrm{T}}\ \tilde{Z}_{jk}(Y_{ik})\}],$$

and we estimate γ by the solution to the equation

$$\sum_{i=1}^n \hat{E}[\partial \log\{f(b_i;\gamma)\}/\partial\gamma] = 0.$$

The conditional distribution of ξ_{ik} given b_i and the observed data is proportional to

$$\xi_{ik}^{\Delta_{ik}} \exp\left[-\xi_{ik} \int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}}\ Z_{ik}(t) + b^{\mathrm{T}}\ \tilde{Z}_{ik}(t)\}\ \mathrm{d}\Lambda_k(t)\right] \phi_k(\xi_{ik}).$$

Thus, the conditional expectation of ξ_{ik} given b_i and the observed data is equal to

$$\frac{\int_{\xi_{ik}} \xi_{ik} \xi_{ik}^{\Delta_{ik}} \exp\left[-\xi_{ik} \int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}}\ Z_{ik}(t) + b^{\mathrm{T}}\ \tilde{Z}_{ik}(t)\}\ \mathrm{d}\Lambda_k(t)\right] \phi_k(\xi_{ik})\ \mathrm{d}\xi_{ik}}{\int_{\xi_{ik}} \xi_{ik}^{\Delta_{ik}} \exp\left[-\xi_{ik} \int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}}\ Z_{ik}(t) + b^{\mathrm{T}}\ \tilde{Z}_{ik}(t)\}\ \mathrm{d}\Lambda_k(t)\right] \phi_k(\xi_{ik})\ \mathrm{d}\xi_{ik}}$$

$$= \begin{cases} G'_k\left[\int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}} Z_{ik}(t) + b^{\mathrm{T}} \tilde{Z}_{ik}(t)\} \mathrm{d}\Lambda_k(t)\right] & \text{if } \Delta_{ik} = 0, \\ -\dfrac{G''_k\left[\int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}} Z_{ik}(t) + b^{\mathrm{T}} \tilde{Z}_{ik}(t)\} \mathrm{d}\Lambda_k(t)\right]}{G'_k\left[\int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}} Z_{ik}(t) + b^{\mathrm{T}} \tilde{Z}_{ik}(t)\} \mathrm{d}\Lambda_k(t)\right]} + G'_k\left[\int_0^{Y_{ik}} \exp\{\beta^{\mathrm{T}} Z_{ik}(t) + b^{\mathrm{T}} \tilde{Z}_{ik}(t)\} \mathrm{d}\Lambda_k(t)\right] & \text{if } \Delta_{ik} = 1. \end{cases}$$

It follows that

$$\hat{E}[\xi_{ik}\exp\{\beta^{\mathrm{T}} Z_{ik}(t)+b_i^{\mathrm{T}}\tilde{Z}_{ik}(t)\}]=\hat{E}\left\{\left(-\Delta_{ik}\frac{G_k''\left[\int_0^{Y_{ik}}\exp\{\beta^{\mathrm{T}} Z_{ik}(t)+b_i^{\mathrm{T}}\tilde{Z}_{ik}(t)\}\,\mathrm{d}\Lambda_k(t)\right]}{G_k'\left[\int_0^{Y_{ik}}\exp\{\beta^{\mathrm{T}} Z_{ik}(t)+b_i^{\mathrm{T}}\tilde{Z}_{ik}(t)\}\,\mathrm{d}\Lambda_k(t)\right]}\right.\right.$$
$$\left.\left.+G_k'\left[\int_0^{Y_{ik}}\exp\{\beta^{\mathrm{T}} Z_{ik}(t)+b_i^{\mathrm{T}}\tilde{Z}_{ik}(t)\}\,\mathrm{d}\Lambda_k(t)\right]\right)\exp\{\beta^{\mathrm{T}} Z_{ik}(t)+b_i^{\mathrm{T}}\tilde{Z}_{ik}(t)\}\right\},$$

which is an integration over b_j only. Conditional on the data, the density of b_i is proportional to

$$\prod_{k=1}^{K}\left\{\exp\{\Delta_{ik}b^{\mathrm{T}}\tilde{Z}_{ik}(Y_{ik})\}\,G_k'\left[\int_0^{Y_{ik}}\exp\{\beta^{\mathrm{T}} Z_{ik}(s)+b^{\mathrm{T}}\tilde{Z}_{ik}(s)\}\,\mathrm{d}\Lambda_k(s)\right]^{\Delta_{ik}}\right.$$
$$\left.\times\exp\left(-G_k\left[\int_0^{Y_{ik}}\exp\{\beta^{\mathrm{T}} Z_{ik}(s)+b^{\mathrm{T}}\tilde{Z}_{ik}(s)\}\,\mathrm{d}\Lambda_k(s)\right]\right)\right\}f(b;\gamma),$$

so the conditional expectation of any function of b can be calculated via high order numerical approximations, such as the high order Gaussian quadrature approximation, the Laplace approximation or Monte Carlo approximations.

On convergence of the algorithm, the Louis (1982) formula is used to calculate the observed information matrix for the parametric and non-parametric components, the latter consisting of the estimated jump sizes in the Λ_ks.

A.2. Recursive formulae

We first consider transformation models without random effects for survival data. Suppose that $\Psi(\mathcal{O}_i;\theta,\Lambda)$ depends on Λ only through $\Lambda(Y_i)$, where Y_i is the observation time for the ith subject. This condition holds if, for example, the covariates are time invariant. We wish to determine the profile likelihood function for θ, i.e. to find the value of Λ that maximizes the objective function for fixed θ. Let $t_1<\ldots<t_m$ be the ordered distinct time points where failures are observed, and let $d_1,\ldots,d_m$ be the jump sizes of Λ at these time points. The likelihood equation that d_k should satisfy is given by

$$0=\frac{1}{d_k}+\sum_{j=1}^{n}I(Y_j\geqslant t_k)\,\nabla_{\Lambda(Y_j)}\log[\Psi\{\mathcal{O}_j;\theta,\Lambda(Y_j)\}],$$

where $\nabla_x g(x,y)=\partial g(x,y)/\partial x$. It follows that

$$\frac{1}{d_{k+1}}=\frac{1}{d_k}+\sum_{j=1}^{n}I(t_k\leqslant Y_j<t_{k+1})\,\nabla_{\Lambda(Y_j)}\log\left\{\Psi\left(\mathcal{O}_j;\theta,\sum_{l=1}^{k}d_l\right)\right\}.$$

This gives a forward recursive formula for calculating the d_k starting from d_1. We can also obtain a backward recursive formula by reparameterizing $\Lambda(x)$ as $\alpha\,F(x)$ with $\alpha=\Lambda(\tau)$ and $F(x)$ a distribution function in $[0,\tau]$. Abusing notation, we write $\Psi(\mathcal{O}_i;\theta,F)$ in which θ now contains α. Since the jump sizes of F add up to 1, the likelihood score equation for the jump size of F at t_{k+1}, which is still denoted as d_{k+1}, satisfies

$$\frac{1}{d_k}=\frac{1}{d_{k+1}}-\sum_{j=1}^{n}I(t_k\leqslant Y_j<t_{k+1})\,\nabla_{\Lambda(Y_j)}\log\left\{\Psi\left(\mathcal{O}_j;\theta,1-\sum_{l=k+1}^{m}d_l\right)\right\}.$$

This is a backward recursive formula for calculating the d_k from d_m. There is one additional constraint: $\Sigma_k\,d_k=1$. It is straightforward to extend the recursive formulae to recurrent events, the only difference being that the summation over individuals is replaced by the double summation over individuals and over events within individuals. For transformation models with random effects, the recursive formulae can be used in the M-step of the EM algorithm.

Appendix B: Technical details

In this appendix, we establish the asymptotic properties of the NPMLEs. A more thorough treatment is given in Zeng and Lin (2007). We first present a general asymptotic theory. We impose the following conditions.

(a) The parameter value θ_0 lies in the interior of a compact set Θ, and Λ_{0k} is continuously differentiable in $[0,\tau]$ with $\Lambda'_{0k}(t)>0$, $k=1,\ldots,K$ (condition (C1)).

(b) With probability 1, $P[\inf_{s\in[0,t]}\{R_{ik\cdot}(s)\}\geqslant 1|Z_{ikl}, l=1,\ldots,n_{ik}]>\delta_0>0$ for all $t\in[0,\tau]$, where $R_{ik\cdot}(t)=\Sigma_{l=1}^{n_{ik}} R_{ikl}(t)$ (condition (C2)).

(c) There is a constant $c_1>0$ and a random variable $r_1(\mathcal{O}_i)>0$ such that $E[\log\{r_1(\mathcal{O}_i)\}]<\infty$ and, for any $\theta\in\Theta$ and any finite $\Lambda_1,\ldots,\Lambda_K$,

$$\Psi(\mathcal{O}_i;\theta,\mathcal{A})\leqslant r_1(\mathcal{O}_i)\prod_{k=1}^{K}\prod_{t\leqslant\tau}\left\{1+\int_0^t R_{ik\cdot}(t)\,\mathrm{d}\Lambda_k(t)\right\}^{-\mathrm{d}N^*_{ik\cdot}(t)}\left\{1+\int_0^\tau R_{ik\cdot}(t)\,\mathrm{d}\Lambda_k(t)\right\}^{-c_1}$$

almost surely, where $N^*_{ik\cdot}(t)=\Sigma_{l=1}^{n_{ik}}N^*_{ikl}(t)$. In addition, for any constant c_2,

$$\inf\{\Psi(\mathcal{O}_i;\theta,\mathcal{A}):\|\Lambda_1\|_{V[0,\tau]}\leqslant c_2,\ldots,\|\Lambda_K\|_{V[0,\tau]}\leqslant c_2,\theta\in\Theta\}>r_2(\mathcal{O}_i)>0,$$

where $\|h\|_{V[0,\tau]}$ is the total variation of $h(\cdot)$ in $[0,\tau]$, and $r_2(\mathcal{O}_i)$ is a random variable with $E\{r_2(\mathcal{O}_i)^6\}<\infty$ and $E[\log\{r_2(\mathcal{O}_i)\}]<\infty$ (condition (C3)).

(d) For any $(\theta^{(1)},\theta^{(2)})\in\Theta$, and $(\Lambda_1^{(1)},\Lambda_1^{(2)}),\ldots,(\Lambda_K^{(1)},\Lambda_K^{(2)}),(H_1^{(1)},H_1^{(2)}),\ldots,(H_K^{(1)},H_K^{(2)})$ with uniformly bounded total variations, there is a function $\mathcal{F}(\mathcal{O}_i)$ in $L_2(P)$ such that

$$\begin{aligned}&|\Psi(\mathcal{O}_i;\theta^{(1)},\mathcal{A}^{(1)})-\Psi(\mathcal{O}_i;\theta^{(2)},\mathcal{A}^{(2)})|+|\dot\Psi_\theta(\mathcal{O}_i;\theta^{(1)},\mathcal{A}^{(1)})-\dot\Psi_\theta(\mathcal{O}_i;\theta^{(2)},\mathcal{A}^{(2)})|\\&\quad+\sum_{k=1}^{K}|\dot\Psi_k(\mathcal{O}_i;\theta^{(1)},\mathcal{A}^{(1)})[H_k^{(1)}]-\dot\Psi_k(\mathcal{O}_i;\theta^{(2)},\mathcal{A}^{(2)})[H_k^{(2)}]|\\&\qquad\leqslant\mathcal{F}(\mathcal{O}_i)\left[|\theta^{(1)}-\theta^{(2)}|+\sum_{k=1}^{K}\left\{\int_0^\tau|\Lambda_k^{(1)}(s)-\Lambda_k^{(2)}(s)|\,\mathrm{d}N_{ik\cdot}+\int_0^\tau|\Lambda_k^{(1)}(s)-\Lambda_k^{(2)}(s)|\,\mathrm{d}s\right\}\right.\\&\qquad\quad\left.+\sum_{k=1}^{K}\left\{\int_0^\tau|H_k^{(1)}(s)-H_k^{(2)}(s)|\,\mathrm{d}N_{ik\cdot}+\int_0^\tau|H_k^{(1)}(s)-H_k^{(2)}(s)|\,\mathrm{d}s\right\}\right],\end{aligned}$$

where $\dot\Psi_\theta$ is the derivative of $\Psi(\mathcal{O}_i;\theta,\mathcal{A})$ with respect to θ, and $\dot\Psi_k[H_k]$ is the derivative of $\Psi(\mathcal{O}_i;\theta,\mathcal{A})$ along the path $(\Lambda_k+\varepsilon H_k)$ (condition (C4)).

(e) If

$$\prod_{k=1}^{K}\prod_{l=1}^{n_{ik}}\prod_{t\leqslant\tau}\lambda_k^*(t)^{R_{ikl}(t)\mathrm{d}N^*_{ikl}(t)}\,\Psi(\mathcal{O}_i;\theta^*,\mathcal{A}^*)=\prod_{k=1}^{K}\prod_{l=1}^{n_{ik}}\prod_{t\leqslant\tau}\lambda_{0k}(t)^{R_{ikl}(t)\mathrm{d}N^*_{ikl}(t)}\,\Psi(\mathcal{O}_i;\theta_0,\mathcal{A}_0)$$

almost surely, then $\theta^*=\theta_0$ and $\Lambda_k^*(t)=\Lambda_{0k}(t)$ for $t\in[0,\tau]$, $k=1,\ldots,K$ (condition (C5); first identifiability condition).

(f) There are functions $\zeta_{0k}(s;\theta_0,\mathcal{A}_0)\in\mathrm{BV}[0,\tau]$, $k=1,\ldots,K$, and a matrix $\zeta_{0\theta}(\theta_0,\mathcal{A}_0)$ such that

$$\begin{aligned}&\left|E\left\{\frac{\dot\Psi_\theta(\mathcal{O}_i;\theta,\mathcal{A})}{\Psi(\mathcal{O}_i;\theta,\mathcal{A})}-\frac{\dot\Psi_\theta(\mathcal{O}_i;\theta_0,\mathcal{A}_0)}{\Psi(\mathcal{O}_i;\theta_0,\mathcal{A}_0)}\right\}-\zeta_{0\theta}(\theta_0,\mathcal{A}_0)^{\mathrm{T}}(\theta-\theta_0)-\sum_{k=1}^{K}\int_0^\tau\zeta_{0k}(s;\theta_0,\mathcal{A}_0)\,\mathrm{d}(\Lambda_k-\Lambda_{0k})\right|\\&\quad=o\left(|\theta-\theta_0|+\sum_{k=1}^{K}\|\Lambda_k-\Lambda_{0k}\|_{V[0,\tau]}\right),\end{aligned}$$

where $\mathrm{BV}[0,\tau]$ denotes the space of functions with bounded total variations in $[0,\tau]$. In addition, for $k=1,\ldots,K$,

$$\begin{aligned}&\sum_{k=1}^{K}\sup_{s\in[0,\tau]}|\{\eta_{0k}(s;\theta,\mathcal{A})-\eta_{0k}(s;\theta_0,\mathcal{A}_0)\}-\eta_{0k\theta}(s;\theta_0,\mathcal{A}_0)^{\mathrm{T}}(\theta-\theta_0)\\&\quad-\int_0^\tau\sum_{m=1}^{K}\eta_{0km}(s,t;\theta_0,\mathcal{A}_0)\,\mathrm{d}(\Lambda_m-\Lambda_{0m})(t)|\\&\qquad\qquad=o\left(|\theta-\theta_0|+\sum_{k=1}^{K}\|\Lambda_k-\Lambda_{0k}\|_{V[0,\tau]}\right),\end{aligned}$$

where $\eta_{0k}(s;\theta,\mathcal{A})$ is a bounded function such that

$$E\{\Psi^{-1}(\mathcal{O}_i;\theta,\mathcal{A})\ \dot{\Psi}_k(\mathcal{O}_i;\theta,\mathcal{A})[H_k]\} = \int_0^\tau \eta_{0k}(s;\theta,\mathcal{A})\ \mathrm{d}H_k(s),$$

η_{0km} is a bounded bivariate function and $\eta_{0k\theta}$ is a d-dimensional bounded function. Furthermore, there is a constant c_3 such that

$$|\eta_{0km}(s,t_1;\theta_0,\mathcal{A}_0) - \eta_{0km}(s,t_2;\theta_0,\mathcal{A}_0)| \leqslant c_3|t_1 - t_2|$$

for any $s \in [0,\tau]$ and any $t_1, t_2 \in [0,\tau]$ (condition (C6)).

(g) If, with probability 1,

$$\sum_{k=1}^{K}\sum_{l=1}^{n_{ik}} \int h_k(t)\ R_{ikl}(t)\ \mathrm{d}N^*_{ikl}(t) + \frac{\dot{\Psi}_\theta(\mathcal{O}_i;\theta_0,\mathcal{A}_0)^{\mathrm{T}} v + \sum_{k=1}^{K} \dot{\Psi}_k(\mathcal{O}_i;\theta_0,\mathcal{A}_0)\left[\int h_k\ \mathrm{d}\Lambda_{0k}\right]}{\Psi(\mathcal{O}_i;\theta_0,\mathcal{A}_0)} = 0$$

for some constant vector $v \in R^d$ and $h_k \in \mathrm{BV}[0,\tau]$, $k = 1,\ldots,K$, then $v = 0$ and $h_k = 0$ for $k = 1,\ldots,K$ (condition (C7); second identifiability condition).

(h) There is a neighbourhood of $(\theta_0,\mathcal{A}_0)$ such that, for $(\theta,\mathcal{A})$ in this neighbourhood, the first and second derivatives of $\Psi(\mathcal{O}_i;\theta,\mathcal{A})$ with respect to θ and along the path $\Lambda_k + \varepsilon H_k$ with respect to ε satisfy the inequality in condition (C4) (condition (C8)).

Theorems 1 and 2 below state the consistency, weak convergence and asymptotic efficiency of the NPMLEs, whereas theorems 3 and 4 justify the use of the observed information matrix and profile likelihood method in the variance–covariance estimation.

Theorem 1. Under conditions (C1)–(C5),

$$|\hat{\theta} - \theta_0| + \sum_{k=1}^{K} \sup_{t\in[0,\tau]} |\hat{\Lambda}_k(t) - \Lambda_{0k}(t)|$$

converges to 0 almost surely.

Theorem 2. Under conditions (C1)–(C7), $n^{1/2}(\hat{\theta}-\theta_0, \hat{\mathcal{A}} - \mathcal{A}_0)$ converges weakly to a zero-mean Gaussian process in $R^d \times l^\infty(\mathcal{Q}^K)$, where $\mathcal{Q} = \{h(t) : \|h(t)\|_{V[0,\tau]} \leqslant 1\}$. Furthermore, the limiting covariance matrix of $n^{1/2}(\hat{\theta}-\theta_0)$ attains the semiparametric efficiency bound.

Theorem 3. Under conditions (C1)–(C8), $n(v^{\mathrm{T}}, \mathbf{h}_1^{\mathrm{T}},\ldots,\mathbf{h}_K^{\mathrm{T}})\ \mathcal{I}_n^{-1}(v^{\mathrm{T}}, \mathbf{h}_1^{\mathrm{T}},\ldots,\mathbf{h}_K^{\mathrm{T}})^{\mathrm{T}}$ converges in probability to the asymptotic variance of

$$n^{1/2}\left\{v^{\mathrm{T}}(\hat{\theta}-\theta_0) + \sum_{k=1}^{K} \int h_k\ \mathrm{d}(\hat{\Lambda}_k - \Lambda_{0k})\right\},$$

where $\mathbf{h}_k$ is the vector consisting of the values of $h_k(\cdot)$ at the observed failure times and $\mathcal{I}_n$ is the negative Hessian matrix of the log-likelihood function with respect to $\hat{\theta}$ and the jump sizes of $(\hat{\Lambda}_1,\ldots,\hat{\Lambda}_K)$.

Theorem 4. Let $\mathrm{pl}_n(\theta)$ be the profile log-likelihood function for θ, and assume that conditions (C1)–(C8) hold. For any $\varepsilon_n = O_p(n^{-1/2})$ and any vector v, $-\{\mathrm{pl}_n(\hat{\theta}+\varepsilon_n v) - 2\ \mathrm{pl}_n(\hat{\theta}) + \mathrm{pl}_n(\hat{\theta}-\varepsilon_n)\}/n\varepsilon_n^2$ converges in probability to $v^{\mathrm{T}}\Sigma^{-1}v$, where Σ is the asymptotic covariance matrix of $n^{1/2}(\hat{\theta}-\theta_0)$.

Theorems 1–4 are proved in Zeng and Lin (2007). To establish the desired asymptotic results for a specific problem, all we need to do is to determine a set of conditions under which regularity conditions (C1)–(C8) are satisfied. As an illustration, we consider the transformation models with random effects for dependent failure times that were described in Section 2.2. We assume the following conditions.

(a) The parameter value $(\beta_0^{\mathrm{T}}, \gamma_0^{\mathrm{T}})^{\mathrm{T}}$ belongs to the interior of a compact set Θ in R^d, and $\Lambda'_{0k}(t) > 0$ for all $t \in [0,\tau]$, $k = 1,\ldots,K$ (condition (D1)).

(b) With probability 1, $Z_{ikl}(\cdot)$ and $\tilde{Z}_{ikl}(\cdot)$ are left continuous in $[0,\tau]$ with uniformly bounded left derivatives (condition (D2)).

(c) With probability 1, $P(C_{ikl} \geqslant \tau | Z_{ikl}) > \delta_0 > 0$ for some constant δ_0 (condition (D3)).

(d) With probability 1, n_{ik} is bounded by some integer n_0. In addition, $E\{N_{ik\cdot}(\tau)\} < \infty$ (condition (D4)).

(e) For $k=1,\ldots,K$, $G_k(x)$ is four times differentiable such that $G_k(0)=0$, $G'_k(x)>0$, and, for any integer $m\geqslant 0$ and any sequence $0<x_1<\ldots<x_m\leqslant y$,

$$\prod_{l=1}^{m}\{(1+x_l)\ G'_k(x_l)\}\ \exp\{-G_k(y)\}\leqslant \mu_{0k}^{m}(1+y)^{-\kappa_{0k}}$$

for some constants μ_{0k} and $\kappa_{0k}>0$. In addition, there is a constant ρ_{0k} such that

$$\sup_x\left\{\frac{|G''_k(x)|+|G^{(3)}(x)|+|G^{(4)}(x)|}{G'(x)(1+x)^{\rho_{0k}}}\right\}<\infty$$

(condition (D5)).

(f) For any constant $a_1>0$,

$$\sup_{\gamma}\left\{E\left(\int_b \exp[a_1\{N^*_{ik\cdot}(\tau)+1\}|b|]\ f(b;\gamma)\ \mathrm{d}b\right)\right\}<\infty,$$

and there is a constant $a_2>0$ such that, for any γ,

$$\left|\frac{\dot f_{\gamma}(b;\gamma)}{f(b;\gamma)}\right|+\left|\frac{\ddot f_{\gamma}(b;\gamma)}{f(b;\gamma)}\right|+\left|\frac{f^{(3)}_{\gamma}(b;\gamma)}{f(b;\gamma)}\right|\leqslant O(1)\exp\{a_2(1+|b|)\}$$

(condition (D6)).

(g) If there are $c(t)$ and v such that $c(t)+v^{\mathrm{T}}\ Z_{ikl}(t)=0$ with probability 1 for $k=1,\ldots,K$ and $l=1,\ldots,n_{ik}$, then $c(t)=0$ and $v=0$. In addition, there is some $t\in[0,\tau]$ such that $\{\tilde Z_{ikl}(t), k=1,\ldots,K, l=1,\ldots,n_{ik}\}$ spans the whole space of b (condition (D7)).

(h) $f(b;\gamma)=f(b;\gamma_0)$ if and only if $\gamma=\gamma_0$; if $v^{\mathrm{T}}\ f'(b;\gamma_0)=0$, then $v=0$ (condition (D8)).

We wish to show that conditions (D1)–(D8) imply conditions (C1)–(C8). Conditions (C1) and (C2) follow naturally from conditions (D1)–(D4). Tedious algebraic manipulations show that conditions (C5) and (C7) hold under conditions (D7) and (D8). Note that

$$\Psi(\mathcal{O}_i;\theta,\mathcal{A})=\int_b\prod_{k=1}^{K}\prod_{l=1}^{n_{ik}}\Omega_{ikl}(b;\beta,\Lambda_k)\ f(b;\gamma)\ \mathrm{d}\mu(b),$$

where

$$\Omega_{ikl}(b;\beta,\Lambda_k)=\prod_{t\leqslant\tau}[R_{ikl}(t)\exp\{\beta^{\mathrm{T}}\ Z_{ikl}(t)+b^{\mathrm{T}}\ \tilde Z_{ikl}(t)\}G'_k\{q_{ikl}(t)\}]^{\mathrm{d}N^*_{ikl}(t)}\ \exp[-G_k\{q_{ikl}(\tau)\}],$$

and

$$q_{ikl}(t)=\int_0^t R_{ikl}(s)\exp\{\beta^{\mathrm{T}}\ Z_{ikl}(s)+b^{\mathrm{T}}\ \tilde Z_{ikl}(s)\}\,\mathrm{d}\Lambda_k(s).$$

If $|b|$ and $\|\Lambda_k\|_{V[0,\tau]}$ are bounded, then $\Omega_{ikl}(b;\beta,\Lambda_k)\geqslant\exp\{O(1)\ N^*_{ikl}(\tau)\}$. Thus, $\Psi(\mathcal{O}_i;\theta,\mathcal{A})$ is bounded from below by $\exp\{O(1)\ N^*_{ikl}(\tau)\}$, so the second half of condition (C3) holds. It follows from condition (D5) that

$$\Omega_{ikl}(b;\beta,\Lambda_k)\leqslant O(1)\prod_{t\leqslant\tau}[R_{ikl}(t)\exp\{b^{\mathrm{T}}\ \tilde Z_{ikl}(t)\}]^{\mathrm{d}N^*_{ikl}(t)}\mu_{0k}^{N^*_{ikl}(\tau)}\prod_{t\leqslant\tau}\{1+q_{ikl}(t)\}^{-\mathrm{d}N^*_{ikl}(t)}\{1+q_{ikl}(\tau)\}^{-\kappa_{0k}}.$$

Since $\exp\{\beta^{\mathrm{T}}\ Z_{ikl}(s)+b^{\mathrm{T}}\ \tilde Z_{ikl}(s)\}\geqslant\exp\{-O(1+|b|)\}$, we have

$$1+q_{ikl}(t)\geqslant\exp\{-O(1+|b|)\}\left\{1+\int_0^t R_{ik\cdot}(s)\ \mathrm{d}\Lambda_k(s)\right\},$$

so

$$\Omega_{ikl}(b;\beta,\Lambda_k)\leqslant O(1)\mu_{0k}^{N^*_{ikl}(\tau)}\exp[O\{1+N^*_{ikl}(\tau)\}|b|]\prod_{t\leqslant\tau}\left\{1+\int_0^t R_{ik\cdot}(s)\,\mathrm{d}\Lambda_k(s)\right\}^{-\mathrm{d}N^*_{ikl}(t)}$$
$$\times\left\{1+\int_0^\tau R_{ikl}(s)\,\mathrm{d}\Lambda_k(s)\right\}^{-\kappa_{0k}}.$$

Thus, the first half of condition (C3) holds as well.

Under condition (D5),

$$|\Omega_{ikl}(b;\beta,\Lambda_k)|\leqslant\exp[O\{1+N^*_{ikl}(\tau)\}|b|],$$
$$|\partial\Omega_{ikl}(b;\beta,\Lambda_k)/\partial\beta|\leqslant\exp[O\{1+N^*_{ikl}(\tau)\}(1+|b|)],$$
$$|\partial\Omega_{ikl}(b;\beta,\Lambda_k)[H_k]/\partial\Lambda_k|\leqslant\exp[O\{1+N^*_{ikl}(\tau)\}(1+|b|)].$$

By the mean value theorem,

$$|\Omega_{ikl}(b;\beta^{(1)},\Lambda_k)-\Omega_{ikl}(b;\beta^{(2)},\Lambda_k)|\leqslant\exp[O\{1+N^*_{ikl}(\tau)\}|b|]|\beta^{(1)}-\beta^{(2)}|,$$
$$|\Omega_{ikl}(b;\beta,\Lambda_k^{(1)})-\Omega_{ikl}(b;\beta,\Lambda_k^{(2)})|\leqslant\exp[O\{1+N^*_{ikl}(\tau)\}(1+|b|)]$$
$$\times\left\{\int R_{ikl}(t)|\Lambda_k^{(1)}(t)-\Lambda_k^{(2)}(t)|\,\mathrm{d}N^*_{ikl}(t)+\int_0^\tau|\Lambda_k^{(1)}(s)-\Lambda_k^{(2)}(s)|\,\mathrm{d}s\right\}.$$

It then follows from condition (D6) that $|\Psi(\mathcal{O}_i;\theta^{(1)},\mathcal{A}^{(1)})-\Psi(\mathcal{O}_i;\theta^{(2)},\mathcal{A}^{(2)})|$ is bounded by the right-hand side of the inequality in condition (C4). The same arguments yield the bounds for the other two terms in condition (C4). The verification of condition (C8) is similar to that of condition (C4), relying on the explicit expressions of $\ddot{\Psi}_{\theta\theta}(\mathcal{O}_i;\theta,\mathcal{A})$ and the first and second derivatives of $\Psi(\mathcal{O}_i;\theta,\mathcal{A}_0+\varepsilon\mathcal{H})$ with respect to ε.

To verify condition (C6), we calculate that

$$\eta_{0k}(s;\theta,\mathcal{A})=E\left(\int_b\frac{\prod_{m=1}^K\prod_{l=1}^{n_{im}}\Omega_{iml}(b;\beta,\Lambda_m)f(b;\gamma)}{\int_b\prod_{m=1}^K\prod_{l=1}^{n_{im}}\Omega_{iml}(b;\beta,\Lambda_m)f(b;\gamma)\,\mathrm{d}b}\left[\int_{t\geqslant s}\frac{G_k''\{q_{ikl}(t)\}}{G_k'\{q_{ikl}(t)\}}\mathrm{d}N^*_{ikl}(t)-G_k'\{q_{ikl}(\tau)\}\right]\right.$$
$$\left.\times R_{ikl}(s)\exp\{\beta^{\mathrm{T}}Z_{ikl}(s)+b^{\mathrm{T}}\tilde{Z}_{ikl}(s)\}\,\mathrm{d}b\right).$$

For $(\theta,\mathcal{A})$ in a neighbourhood of $(\theta_0,\mathcal{A}_0)$,

$$\left|\eta_{0k}(s;\theta,\mathcal{A})-\eta_{0k}(s;\theta_0,\mathcal{A}_0)-\frac{\partial}{\partial\theta}\eta_{0k}(s;\theta_0,\mathcal{A}_0)^{\mathrm{T}}(\theta-\theta_0)-\sum_{m=1}^K\frac{\partial\eta_{0k}}{\partial\Lambda_m}(s;\theta_0,\mathcal{A}_0)[\Lambda_m-\Lambda_{0m}]\right|$$
$$=o\left(|\theta-\theta_0|+\sum_{m=1}^K\|\Lambda_m-\Lambda_{0m}\|_{V[0,\tau]}\right).$$

Thus, for the second equation in condition (C6), $\eta_{0km}(s,t;\theta_0,\mathcal{A}_0)$ is obtained from the derivative of η_{0k} with respect to Λ_m along the direction $\Lambda_m-\Lambda_{0m}$, and $\eta_{0k\theta}$ is the derivative of η_{0k} with respect to θ. Likewise, we can obtain the first equation in condition (C6). It is straightforward to verify the Lipschitz continuity of η_{0km}.

The asymptotic properties for the other models in this paper are verified in Zeng and Lin (2007).

References

Akaike, H. (1985) Prediction and entropy. In *A Celebration of Statistics* (eds A. C. Atkinson and S. E. Fienberg), pp. 1–24. New York: Springer.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical Models based on Counting Processes*. New York: Springer.

Andersen, P. K. and Gill, R. D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1100–1120.

Andersen, P. K., Klein, J. P., Knudsen, K. M. and Tabanera y Placios, R. (1997) Estimation of Cox's regression model with shared gamma frailties. *Biometrics*, **53**, 1475–1484.

Bagdonavicius, V., Hafdi, M. A. and Nikulin, M. (2004) Analysis of survival data with cross-effects of survival functions. *Biostatistics*, **5**, 415–425.

Bennett, S. (1983) Analysis of survival data by the proportional odds model. *Statist. Med.*, **2**, 273–277.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pagoda, J. (2000) Exposure stratified case-cohort designs. *Liftime Data Anal.*, **6**, 39–58.

Box, G. E. P. and Cox, D. R. (1982) An analysis of transformations revisited, rebutted. *J. Am. Statist. Ass.*, **77**, 209–210.

Breslow, N. (1972) Discussion on 'Regression models and life-tables' (by D. R. Cox). *J. R. Statist. Soc.* B, **34**, 216–217.

Cai, T., Cheng, S. C. and Wei, L. J. (2002) Semiparametric mixed-effects models for clustered failure time data. *J. Am. Statist. Ass.*, **97**, 514–522.

Chen, H. Y. (2002) Double-semiparametric method for missing covariates in Cox regression models. *J. Am. Statist. Ass.*, **97**, 565–576.

Chen, H. Y. and Little, R. J. A. (1999) Proportional hazards regression with missing covariates. *J. Am. Statist. Ass.*, **94**, 896–908.

Chen, K., Jin, Z. and Ying, Z. (2002) Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**, 659–668.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995) Analysis of transformation models with censored data. *Biometrika*, **82**, 835–845.

Clayton, D. and Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model (with discussion). *J. R. Statist. Soc.* A, **148**, 82–117.

Coleman, T. F. and Li, Y. (1994) On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. *Math. Programng*, **67**, 189–224.

Coleman, T. F. and Li, Y. (1996) An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optimizn*, **6**, 418–445.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.

Dabrowska, D. M. and Doksum, K. A. (1988) Partial likelihood in transformation models with censored data. *Scand. J. Statist.*, **18**, 1–23.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Diao, G. and Lin, D. Y. (2005) A powerful and robust method for mapping quantitative trait loci in general pedigrees. *Am. J. Hum. Genet.*, **77**, 97–111.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*, 2nd edn. Oxford: Oxford University Press.

Farewell, V. T. (1982) The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 1041–1046.

Gilat, A. (2004) *MATLAB: an Introduction with Applications*, 2nd edn. Hoboken: Wiley.

Henderson, R., Diggle, P. and Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **4**, 465–480.

Hogan, J. W. and Laird, N. M. (1997) Model-based approaches to analysing incomplete longitudinal and failure time data. *Statist. Med.*, **16**, 259–272.

Horowitz, J. L. (1998) *Semiparametric Methods in Economics*. New York: Springer.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. New York: Springer.

Hsieh, F. (2001) On heteroscedastic hazards regression models: theory and application. *J. R. Statist. Soc.* B, **63**, 63–79.

Huang, C. Y. and Wang, M. C. (2004) Joint modeling and estimation for recurrent event processes and failure time data. *J. Am. Statist. Ass.*, **99**, 1153–1165.

Huang, J. (1996) Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.*, **24**, 540–568.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001) *Bayesian Survival Analysis*. New York: Springer.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003) Rank-based inference for the accelerated failure time model. *Biometrika*, **90**, 341–353.

Jin, Z., Lin, D. Y. and Ying, Z. (2006) On least-squares regression with censored data. *Biometrika*, **93**, 147–161.

Kalbfleisch, J. D. and Lawless, J. F. (1988) Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med.*, **7**, 147–160.

Kalbfleisch, J. D. and Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edn. Hoboken: Wiley.

Klein, J. (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**, 795–806.

Kosorok, M. R., Lee, B. L. and Fine, J. P. (2004) Robust inference for proportional hazards univariate frailty regression models. *Ann. Statist.*, **32**, 1448–1491.

Kulich, M. and Lin, D. Y. (2004) Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Am. Statist. Ass.*, **99**, 832–844.

Lin, D. Y. (1994) Cox regression analysis of multivariate failure time data: the marginal approach. *Statist. Med.*, **13**, 2233–2247.

Lin, D. Y. and Ying, Z. (2003) Semiparametric regression analysis of longitudinal data with informative drop-outs. *Biostatistics*, **4**, 385–398.

Lin, D. Y. and Zeng, D. (2006) Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *J. Am. Statist. Ass.*, **101**, 89–118.

Liu, L., Wolfe, R. A. and Huang, X. (2004) Shared frailty models for recurrent events and a terminal event. *Biometrics*, **60**, 747–756.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.

Lu, W. and Ying, Z. (2004) On semiparametric transformation cure models. *Biometrika*, **91**, 331–343.

Murphy, S. A. (1994) Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.*, **22**, 712–731.

Murphy, S. A. (1995) Asymptotic theory for the frailty model. *Ann. Statist.*, **23**, 182–198.

Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997) Maximal likelihood estimation in the proportional odds model. *J. Am. Statist. Ass.*, **92**, 968–976.

Murphy, S. A. and van der Vaart, A. W. (2000) On profile likelihood. *J. Am. Statist. Ass.*, **95**, 449–485.

Nan, B., Emond, M. and Wellner, J. A. (2004) Information bounds for Cox regression models with missing data. *Ann. Statist.*, **32**, 723–753.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, **19**, 25–43.

Oakes, D. (1989) Bivariate survival models induced by frailties. *J. Am. Statist. Ass.*, **84**, 487–493.

Oakes, D. (1991) Frailty models for multiple event times. In *Survival Analysis: State of the Art* (eds J. P. Klein and P. K. Goel), pp. 371–379. Dordrecht: Kluwer.

Parner, E. (1998) Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.*, **26**, 183–214.

Peng, Y. and Dear, K. B. G. (2000) A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**, 237–243.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992) *Numerical Recipes in C*. New York: Cambridge University Press.

Robins, J. M. and Rotnitzky, A. (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues* (eds N. P. Jewell, K. Dietz and V. T. Farewell), pp. 297–331. Boston: Birkhäuser.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression-coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.

Sasieni, P. (1992) Information bounds for the conditional hazard ratio in a nested family of regression models. *J. R. Statist. Soc.* B, **54**, 617–635.

Satterthwaites, F. E. (1946) An approximate distribution of estimates of variance components. *Biometrics*, **2**, 110–114.

Scheike, T. H. and Juul, A. (2004) Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics*, **5**, 193–206.

Scheike, T. H. and Martinussen, T. (2004) Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scand. J. Statist.*, **31**, 283–293.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Self, S. G. and Liang, K. Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Statist. Ass.*, **82**, 605–610.

Stablein, D. M. and Koutrouvelis, I. A. (1985) A two-sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics*, **41**, 643–652.

Sy, J. P. and Taylor, J. M. (2000) Estimation in a Cox proportional hazards cure model. *Biometrics*, **56**, 227–236.

Tsiatis, A. A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statist. Sin.*, **14**, 793–818.

Tsodikov, A. (2003) Semiparametric models: a generalized self-consistency approach. *J. R. Statist. Soc.* B, **65**, 759–774.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.

Wang, M. C., Qin, J. and Chiang, C. (2001) Analyzing recurrent event data with informative censoring. *J. Am. Statist. Ass.*, **96**, 1057–1065.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Statist. Ass.*, **84**, 1065–1073.

Wellner, J. A. and Zhang, Y. (2005) Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Technical Report 488*. Department of Statistics, University of Washington, Seattle.

Wulfsohn, M. S. and Tsiatis, A. A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Xu, J. and Zeger, S. L. (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. *Appl. Statist.*, **50**, 375–387.

Zeng, D. and Lin, D. Y. (2007) A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Technical Report*. University of North Carolina, Chapel Hill.

Zeng, D., Lin, D. Y. and Yin, G. (2005) Maximum likelihood estimation for the proportional odds model with random effects. *J. Am. Statist. Ass.*, **100**, 470–483.

Discussion on the paper by Zeng and Lin

Robin Henderson (*Newcastle University*)

Zeng and Lin have provided a beautifully structured paper whose development in some ways mimics the history of event history methodology over the last 35 years: standard survival and proportional hazards; recurrent events and counting processes; clustered events and frailty; joint modelling for longitudinal and event history data. For discussion I shall concentrate mainly on Zeng and Lin's approach for single-event survival analysis, recognizing that their methods go much further.

The paper starts with the statement that the Cox (1972) proportional hazards model is the corner-stone of modern survival analysis. Although this may not be true for reliability, it is indeed so for biostatistics. Even in 1998 Niels Keiding remarked that '… explicit excuses are now needed to use different models' (Keiding, 1998) and very little has changed since then. The proportionality assumption can and should be challenged but the basic model is so well known and well used that it makes sense to ensure that the standard proportional hazards model is nested within more flexible alternative semiparametric models. In Section 2.1 Zeng and Lin do just that, with equation (4) providing an interesting and potentially very useful extension. This model also incorporates as special cases alternative generalized versions that have been proposed by, for instance, Bagdonavicius and Nikulin (1999), Hseih (2001) and Bagdonavicius *et al.* (2004).

The price to be paid for generality, however, is lack of transparency of the role of covariates. It is difficult to look at an expression like equation (4) and to gain any intuitive impression of exactly how a covariate will influence the hazard. To explore, it is convenient to consider perhaps the simplest case of single-event survival times with two groups, described by a time constant binary covariate Z, with $Z=0$ in the control group and $Z=1$ in the treatment group. Writing $\theta=\exp(\beta)$ and $\phi=\exp(\gamma)$, under the Box–Cox transformations the hazards corresponding to model (4) become

$$\lambda(t|Z)=\begin{cases}\lambda(t)\{1+\Lambda(t)\}^{\rho-1} & Z=0 \text{ (control)},\\ \theta\phi\{1+\theta\,\Lambda(t)\}^{\phi\rho-1}\,\lambda(t) & Z=1 \text{ (treatment)}.\end{cases}$$

For exploration we shall assume that $\lambda(t|Z=0)=1$. We can obtain this in two ways: either by $\rho=1$ and $\lambda(t)=1$, or by $\Lambda(t)=(\rho t+1)^{1/\rho}-1$ for any $\rho>0$. We can take these in turn, starting with $\rho=1$ and $\lambda(t)=1$, which gives

$$\lambda(t|Z)=\begin{cases}1 & Z=0 \text{ (control)},\\ \theta\phi(1+\theta t)^{\phi-1} & Z=1 \text{ (treatment)}.\end{cases}\tag{14}$$

Interpretation in this case is straightforward. The initial hazard in the treatment group is determined by the ratio $\theta\phi$. For $\phi>1$ we have a monotonic increasing hazard (to ∞); for $\phi<1$ monotonic decreasing (to 0). At $\phi=1$ we have a proportional hazards model.

Now taking $\Lambda(t)=(\rho t+1)^{1/\rho}-1$ we have

$$\lambda(t|Z)=\begin{cases}1 & Z=0 \text{ (control)},\\ \theta\phi[1+\theta\{(\rho t+1)^{1/\rho}-1\}]^{\phi\rho-1}(\rho t+1)^{1/\rho-1} & Z=1 \text{ (treatment)}.\end{cases}\tag{15}$$

The initial value is again determined by $\theta\phi$ but otherwise it is difficult to see the role of the three parameters. The treatment hazard tends to ∞ if $\phi>1$, to 0 if $\phi<1$ and to θ^ρ if $\phi=1$. Formally the choice between expressions (14) and (15) is identifiable from data, given a sufficiently large sample. In practice I suspect that it will be difficult and wonder whether the authors have encountered any identifiability problems in their work. Perhaps in the example above it will not matter as the fitted hazard or survival curves under the alternatives may be very close. And this brings us back to interpretation: if the model parameters are difficult to interpret do the authors advocate inspecting fitted curves? More generally, it would be helpful if Zeng and Lin could comment on what real advantages for single-event survival are provided by equation (4) over generalized additive versions of the proportional hazards model (e.g. Sasieni and Winnett (2003)):

$$\lambda(t|Z) = \exp\{\beta(t)\, Z(t)\}\, \lambda(t)$$

or

$$\lambda(t|Z) = \exp\{\sum_j f(t, Z_j)\}\, \lambda(t).$$

Returning to the two-group illustration, another point is that neither expression (14) nor expression (15) can accommodate converging hazards unless $\rho = 0$, which in many ways should be the default model for time constant covariates. Indeed, in discussion of Aalen and Gjessing (2001), Keiding pointed to a paper which is even more venerable than Cox (1972), namely Tetens (1786), where the author took it as almost axiomatic that hazards converge over time and suggested the hazard ratio model

$$\frac{\lambda(t)}{\lambda_0(t)} = \frac{1 + 2\alpha\, S(t)}{1 + \alpha\, S_0(t)}.$$

The second family of transformations, $G(x) = \log(1 + rx)/r$, can, I believe, describe converging hazards, at least for the non-crossing alternative with $\gamma = 0$. In the two-group case the hazards are

$$\lambda(t|Z) = \begin{cases} \lambda(t)/\{1 + r\,\Lambda(t)\} & Z = 0 \text{ (control)}, \\ \theta\lambda(t)/\{1 + r\,\theta\Lambda(t)\} & Z = 1 \text{ (treatment)} \end{cases}$$

and more generally, but still with time constant covariates, the assumed survivor function is

$$S(t|Z) = \left\{ \frac{1}{1 + r\exp(\beta Z)\Lambda(t)} \right\}^{1/r}.$$

This is precisely the marginal survivor function for gamma frailty within a proportional hazards model

$$\Lambda(t|Z, \xi) = \xi \exp(\beta Z)\, \Lambda(t) \qquad \xi \sim \Gamma(1/r, 1/r).$$

I wonder whether the authors have experience of estimating r for single-event survival data. My own experience is that there is usually downward bias in modest sample sizes. In the simulations was r fixed at the true value or estimated?

All of the above is for single-event survival with standard independent right censoring. Of course the paper goes much further and perhaps its main advantages with time will prove to be for the more complex situations that are now being more often considered in applications. Direct maximum likelihood estimation of the jumps in Λ has often been considered infeasible (e.g. Tsodikov (2003)) but the authors have shown this not to be so. What is particularly impressive about the paper is the achievement of realistic computing times for joint modelling of longitudinal and event history data. Zeng and Lin quote a computing time of 35 min at $n = 200$ for random intercept and slope models. Similar models, though without transformations, were considered by me and colleagues Peter Diggle and Angela Dobson and could take many days to fit, meaning that it was realistic only to fit one or a small number of models in any application. Quick computation combined with the availablity of a maximized log-likelihood means straightforward model comparison and proper statistical practice.

It gives me great pleasure to propose the vote of thanks.

Odd O. Aalen (*University of Oslo*)

The paper by Zeng and Lin is an interesting extension of the frailty models in survival and event history analysis, complementing and extending previous work which has for instance been summarized by Hougaard (2000). One feels, however, that they might have gone even further in their frailty formulation. In fact, the transformation model in formula (2) is most easily understood as the result of a frailty factor operating on the Cox proportional hazards model. The authors in fact use this indirectly at the end of Section 3 and in the beginning of Appendix A.1, but they could have given this formulation from the outset. If the intensity for an individual is $U \exp\{\beta^{\mathrm{T}} Z(t)\}\, \lambda(t)$ where U is a frailty variable with Laplace transform $L(x)$, then we obtain essentially formula (2) when the distribution of U is integrated out and we put $G(x) = -\log\{L(x)\}$. Of course, this is just valid when the function $G(x)$ has this kind of representation for some Laplace transform $L(x)$. It is interesting to note that the Box–Cox transformation for $0 \leqslant \rho \leqslant 1$ corresponds to $L(x)$ being the Laplace transform of a power variance family of frailty distributions which played a central role in Hougaard (2000). For $\rho > 1$, it is not clear that there is any frailty representation at all; remarks at the beginning of Appendix A.1 seem to indicate that there should be one in this case as well, and the authors should explain this. We may, however, also use $\rho < 0$; that would give a compound Poisson frailty distribution (Aalen, 1988, 1992; Moger *et al.*, 2004), which is still a subgroup of the power variance family class. In this case there will be a group of individuals with zero risk, but this just means that we have a cure model which would be quite reasonable in many cases.

A basic idea in the paper is to extend the semiparametric principle in the Cox analysis with a non-parametric $\Lambda(t)$, and the rest of the model being parametric. Although the authors elegantly handle the technical problems that are connected to this, one could ask whether the semiparametric principle is reasonable in this case. In the original Cox model, the semiparametric idea yielded a very elegant solution in terms of the partial likelihood. However, this simplicity entirely disappears in the more general context here, and one could ask whether the semiparametric idea is carried too far. With a purely parametric version one would avoid many complications and, anyway, parametric models are highly underused in biostatistical survival analysis.

A nice aspect of the approach by Zeng and Lin is the ability to handle crossing curves. There is in survival analysis too much emphasis on the rather dogmatic assumption of proportional hazards, in spite of the fact that we often observe deviations from this assumption. Professional statisticians would be aware of the mathematical convenience nature of this assumption and how to handle deviations, but other people doing proportional hazards regression, like many medical researchers, might not have a clear view of this limitation and they are not helped by standard statistical software. By the way, crossing of hazard rates could easily be a frailty effect; see for example Aalen (1994).

Another interesting aspect is the general frailty structure for recurrent event models. An alternative to this would be to use a model with dynamic covariates, e.g. according to Fosen *et al.* (2006).

In Section 7 the authors make a rather sweeping statement concerning the use of martingale methods, saying:

> 'The counting process martingale theory ... plays no role in establishing the asymptotic theory for the kind of problem that is considered in this paper .... We have relied heavily on modern empirical process theory, which we believe will be the primary mathematical tool in survival analysis and semiparametric inference more broadly for the foreseeable future.'

It is true that the general model in the paper can be handled by empirical processes, and a solution with martingale central limit theorems does not seem to have been worked out, although such a solution might exist. Clearly, if $\Lambda(t)$ is parametric, martingale theory can be used. It should then be possible to use a sieve approach (Bunea and McKeague, 2005) to approximate the non-parametric $\Lambda(t)$ and to obtain asymptotic results within the martingale framework.

It should be noted, however, that the main point of using martingale theory in survival analysis is not to achieve the asymptotics but to obtain a conceptual underpinning under the statistical approaches. Censoring, for instance, may depend on what has happened previously in the processes. The martingale formulation allows very general assumptions on the censoring mechanisms, which are related to the fundamental martingale concept of optional stopping time.

More generally, the martingale structure is not imposed from the outside but originates in the heart of the processes themselves. This is connected to the 'French probability school' which views stochastic processes in terms of how the past influences the future and the present. A major result here is the Doob–Meyer decomposition by which a semimartingale is decomposed into the sum of a compensator and a martingale. It was first proved by Brémaud in 1972 that this result is precisely what we need to obtain a precise definition of the intensity process (see references in Andersen *et al.* (1993)). The intensity process is a general concept embedding the hazard rate, and at the very heart of modern survival analysis. Before Brémaud's work the definitions of an intensity process were intuitive and had no mathematical precision.

The classical framework of statistics is the assumption of independence in various forms, and this assumption is also necessary for the empirical process approach of Zeng and Lin. The trouble with independence is that it is immediately destroyed once you apply, for instance, a censoring mechanism that is dependent on all the processes under observation. This is not so for the martingale assumption which enjoys fundamental invariance properties expressed by preservation under optional stopping and stochastic integration. Also covariates can depend on the past in complex ways. Moreover, some statistically important quantities are martingales when they are properly normalized; the likelihood and the associated score functions, the Kaplan–Meier and Nelson–Aalen estimators, the empirical transition matrix, the two-sample test statistics and many other important quantities. The introduction of ideas from the French probability school to survival analysis has been presented in several text-books, e.g. Andersen *et al.* (1993) and Martinussen and Scheike (2006). Recent applications to general longitudinal data are given by Martinussen and Scheike (2006), chapter 11, Borgan *et al.* (2007) and Farewell (2006).

There is also a central limit theory for martingales which nicely complements the aspects that were discussed above. There might be limitations in the application of this in some cases, but that is no justification for advocating a general return to the strait-jacket of the independent and identically distributed data world.

Martingale theory has recently revolutionized the field of mathematical finance, and it is high time that the world of statistics also realizes its usefulness. Martingales should be an integrated part of the curriculum of statistics students.

In spite of my critical comments I consider the paper by Zeng and Lin to be of considerable interest. I thank Zeng and Lin for a challenging paper which is also based on a large and impressive effort to make all the technical details work. It gives me great pleasure in congratulating the authors on their paper and in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Daniel Commenges (*University of Bordeaux 2*)
The authors must be congratulated for proposing a general model encompassing multivariate failure time data, frailty models and joint models, for proving asymptotic results for the non-parametric maximum likelihood estimators (NPMLEs) in this model and for proposing maximization algorithms. I have two comments: one on a drawback of the NPMLE; the other on an alternative algorithm.

Although, as shown by the authors, the NPMLE has the advantage of being more efficient than most other estimators it has the drawback of yielding estimators which are not *a priori* in the class of admissible estimators for most applications: the NPMLE of the compensator of a counting process makes jumps whereas we would generally expect that, under the true law, the counting process has an absolutely continuous compensator and admits an intensity. From a descriptive point of view this drawback leads to representing the compensator itself or the survival function only, and not risk functions or transition intensities. This drawback leads to another limitation in that it makes likelihood cross-validation unusable: generally the NPMLE estimate of the risk at the time of event of a removed observation is 0 so the cross-validation criterion takes an infinite value. Thus most often the NPMLE is strongly rejected by a likelihood cross-validation criterion. Likelihood cross-validation 'estimates', up to a constant, the Kullback–Leibler risk, and the NPMLE is not consistent for the Kullback–Leibler risk. An alternative is to use a penalized likelihood yielding smooth compensators and intensities (O'Sullivan, 1988); the advantage of this approach is that likelihood cross-validation may be used for choosing both the structure of the model and the smoothing coefficient, as proposed in Commenges *et al.* (2007). However, deriving the asymptotic properties of the resulting estimator is still an open problem in the general case.

The numerical problem is of course crucial for complex models and the authors investigate several possibilities. I would like to draw attention to an algorithm for maximizing likelihoods, which has already been used by several researchers, and which is based on using the empirical variance of the score in place of the Hessian, thus sparing much computation time; I call it the 'robust variance scoring' algorithm (Commenges *et al.*, 2006). When the function to maximize is the log-likelihood, this algorithm is superior to the BFGS algorithm which is not specific. Even with frailties this algorithm can be used because the observed scores can be obtained by numerical integration from the score of a full problem by using Louis formulae, such as in Hedeker and Gibbons (1994). It would be interesting to try this algorithm on the model that is proposed in this paper.

Torben Martinussen and Thomas H. Scheike (*University of Copenhagen*)
We congratulate the authors on this very interesting and impressive paper. The class of semiparametric transformation models is an appealing class as it accommodates the crossing hazards situation. It has, however, the weakness of not being able to describe time varying covariate effects in a direct interpretable way. Time varying effects are easily estimated by using Aalen's additive hazards model (Aalen, 1980)

$$\lambda(t|Z)=\beta_0(t)+\beta_1^{\mathrm{T}}(t)Z,$$

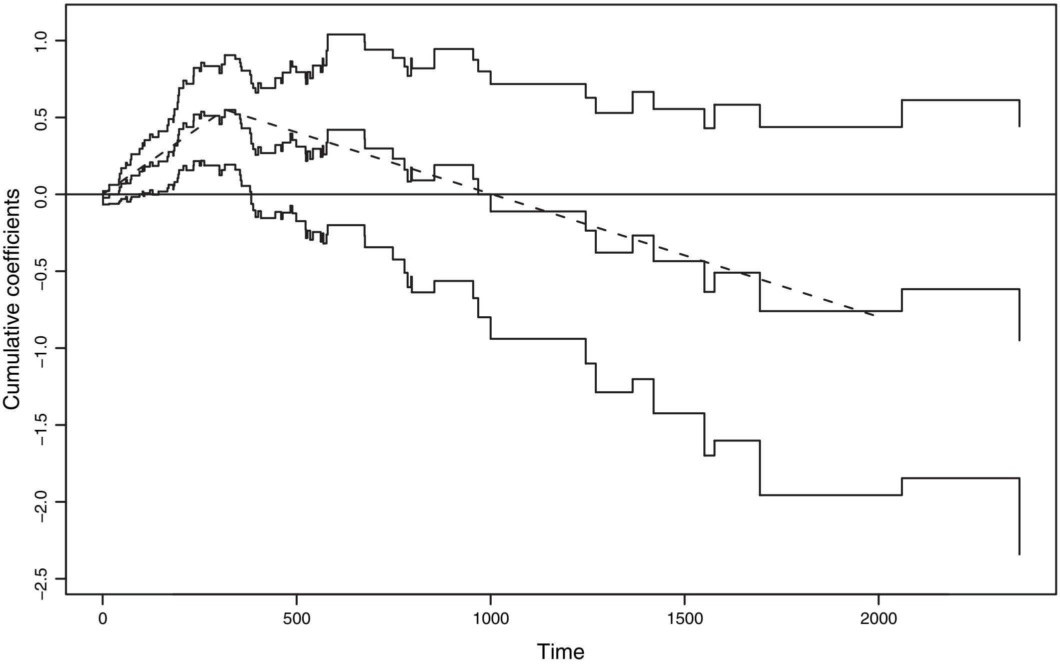

Fig. 5. Gastrointestinal tumour data: effect of combined therapy—Aalen's least squares estimate of $B_1(t)$ with 95% pointwise confidence bands (-------, estimate based on a changepoint model)

where $(\beta_0(t), \beta_1^{\mathrm{T}}(t))$ are unspecified regression functions. We would also like to mention the semiparametric additive risk model of McKeague and Sasieni (1994)

$$\lambda(t|Z) = \beta_0(t) + \beta_1^{\mathrm{T}}(t) Z + \gamma^{\mathrm{T}} X,$$

where some effects are time varying and others are constant (see also Martinussen and Scheike (2006)).

We fitted Aalen's additive hazards model to the gastrointestinal tumour data. The cumulative regression coefficient corresponding to the combined therapy group is depicted in Fig. 5, showing nicely that a time varying effect is indeed the case for these data with a negative effect of the combined treatment in the first 300 days or so, and then apparently an adverse effect thereafter. Another appealing model for these data is the changepoint model (Martinussen and Scheike, 2007)

$$\lambda(t|Z) = \beta_0(t) + Z\{\gamma_1 + \gamma_2 I(t > \theta)\},$$

where Z is a scalar covariate, and γ_1, γ_2 and θ are unknown parameters with the last being the changepoint parameter. For the gastrointestinal tumour data we obtain the estimates $\hat{\theta} = 315$ and $(\hat{\gamma}_1, \hat{\gamma}_2) = (0.00175, -0.00255)$ with the estimated cumulative regression function superimposed on the Aalen estimator in Fig. 5 indicating that this model gives a good fit to these data. The Aalen additive hazards model and its corresponding changepoint model are easily formulated in multivariate covariate settings. We wonder whether it is possible, on the basis of model (4) in the present paper, to pinpoint a time varying effect of a specific covariate in a multiple covariate setting with time constant effect of the remaining covariates, say.

The use of non-parametric maximum likelihood estimators (NPMLEs) and their asymptotic properties in the setting considered are very useful. We wonder how far this can be taken in terms of other classes of models, some of them describing time varying effects of some of the covariates.

The NPMLE behaves sensibly for Aalen's additive hazards model in the situation with only one categorical covariate but seems to break down in the multiple-covariate setting with some of them being continuous. It would also be interesting to clarify whether NPMLEs can be applied to other general models as for example the extended proportional odds model

$$S(t|Z) = [1 + G(t) \exp\{Z^{\mathrm{T}} \beta(t)\}]^{-1}$$

with hazard rate

$$\lambda(t|Z) = S(t|Z) \exp\{Z^{\mathrm{T}} \beta(t)\}\{G'(t) + G(t) Z^{\mathrm{T}} \beta'(t)\},$$

or to a variant of this model, using a first-order approximation of the term $\exp\{Z^{\mathrm{T}} \beta(t)\}$ (Scheike, 2006).

Kani Chen (*Hong Kong University of Science and Technology*) **and Zhiliang Ying** (*Columbia University, New York*)

The pursuit of efficient estimation is always an important problem for nearly all statistical models, especially those of a semiparametric nature. The computational complexity and theoretical justification are main hurdles for obtaining efficient estimators in general semiparametric models. However, the current computational technology, especially the fast developing statistical and mathematical software and the availability of computation capacity, has drastically reduced the computational workload and time. It has made many statistical problems that were previously considered as intractable now readily solvable. Professor Zeng and Professor Lin made this important contribution of computing and justifying efficient estimation for a broad class of semiparametric models with censored data. We congratulate them for such an important development.

Their method seems to be far reaching and has many potential applications and extensions. Heuristically, this method should work if the infinite dimensional parameter, which is typically a function, can be properly discretized so that support of the likelihood function is on a finite dimensional space. The maximization procedure can then take advantage of the available computational algorithms. The idea may date back to the discovery of the empirical distribution as the non-parametric maximum likelihood estimator of the cumulative distribution function. We use the following examples to illustrate further extensions.

Consider the transformation model $H(T_i) = -\beta' Z_i + \varepsilon_i$ for doubly censored data. Let (L_i, U_i) be the censoring variables, and the observations are (Y_i, Z_i, δ_i), where $Y_i = T_i$ and $\delta_i = 1$ if $T_i \in [L_i, U_i]$, $Y_i = U_i$ and $\delta_i = 2$ if $T_i > U_i$, and $Y_i = L_i$ and $\delta_i = 3$ if $T_i < L_i$. The likelihood is

$$L_n(\beta, H) = \prod_{i=1}^{n} \{[\lambda\{\beta' Z_i + H(Y_i)\} h(Y_i)]^{I(\delta_i=1)} \exp[-\Lambda\{\beta' Z_i + H(Y_i)\}\{I(\delta_i = 1) + I(\delta_i = 2)\}] \\ \times (1 - \exp[-\Lambda\{\beta' Z_i + H(Y_i)\}])^{I(\delta_i=3)}\}$$

where $\lambda(\cdot)$ and $\Lambda(\cdot)$ are respectively the known hazard and cumulative hazard functions of ε_i. Restricting $H(\cdot)$ to be step functions with jumps only at the uncensored observations, the likelihood function can be maximized. The maximizers $(\hat{\beta}, \hat{H}(\cdot))$ are consistent and asymptotically normal under suitable conditions.

In this case, there is no backward or forward recursive algorithm for computation, unlike those presented in Chen *et al.* (2002) and this paper for the right-censored data. However, using the function `fmincon` in MATLAB works reasonably well in our simulation studies. For example, we simulated the transformation model with one covariate and with ε following one of the Pareto family of distributions with $r = 0, 0.5, 1$. Note that $r = 0$ and $r = 1$ correspond to the proportional hazards model and the proportional odds model respectively. With censoring percentages that are well over 50% and sample sizes at 100 and 200, we find that the resulting point and variance estimators are virtually unbiased and confidence intervals have coverage probabilities that are close to their nominal levels.

Another type of data is the left-truncated data (T_i, C_i, Z_i) where C_i is the truncation variable. The likelihood is

$$L_n(\beta, H) = \prod_{i=1}^{n} \lambda\{\beta' Z_i + H(T_i)\} h(T_i) \exp[-\Lambda\{\beta' Z_i + H(T_i)\} + \Lambda\{\beta' Z_i + H(C_i)\}].$$

The maximization procedure is similar to that for right-censored data. Certain technical conditions on the distribution of the truncation variable near 0 will be required to ensure proper large sample behaviour of the maximizers.

Alex Tsodikov (*University of Michigan, Ann Arbor*)

I congratulate the authors on this interesting and general paper. I have a few comments on models and the justification for EM algorithms resulting from the quasi-expectation–maximization (QEM) approach of Tsodikov (2003).

Models can be constructed by using a transform $\gamma(x) = \mathrm{QE}(x^u)$, where QE is an operator that is defined so that γ behaves like a probability-generating function $E(x^U)$ of a random variable U up to its first k

moments. The resulting 'artificial' mixed model lends itself naturally to EM-like QEM algorithms. The E-step is represented by using derivatives of the transform γ much like moments of a random variable are represented by using derivatives of its probability-generating function. When the ith derivatives satisfy $(-1)^i \mathrm{d}\gamma^{(i-1)} \exp(-s)/\mathrm{d}s > 0$, $i = 1, \ldots, k$, the resultant algorithm is monotonic in likelihood and QE satisfies the Jensen inequality. The matrix speed of convergence of the algorithm is determined by the fraction of 'missing information' $I_0^{-1} I_{\mathrm{M}}$, where I_0 and I_{M} are 'complete-data' and 'missing data' information matrices respectively that are also expressed through derivatives of the transform.

The authors used an 'artificial' EM construction in their paper to deal with the transformation G that was introduced in equation (2) by suggesting that $\exp(-G)$ be a Laplace transform, and later expressing the E-step by using derivatives of G. It can be shown by using the Bernstein theorem (Feller, 1991) that $\exp(-G)$ is a Laplace transform if and only if the derivative G' satisfies $(-1)^{i+1} G^{(i)} > 0$, $i = 1, 2, \ldots, \infty$. Linking this to QEM, we have $\gamma(x) = \exp[-G\{-\log(x)\}]$. As a QEM the algorithm is valid under the weaker condition of $G' > 0, G'' < 0$ and G'' being an increasing function. The logarithmic transform satisfies the condition for any $r \geqslant 0$, whereas it is necessary that $0 \leqslant \rho < 1$ for the Box–Cox transform. Potentially, G may be represented by a function with discontinuous high order derivatives (splines), in which case $\exp(-G)$ will not be a Laplace transform. However, the algorithm will still converge provided that the weak QEM condition is satisfied. Also, it can be shown that a composition of G-based QEM and an EM dealing with the random effects b will preserve the necessary conditions for monotonic convergence of likelihood values.

Thomas H. Scheike and Torben Martinussen (*University of Copenhagen*)
We are very pleased that the authors take up the theme of random-effects models for survival data where there is clearly much more to be done. One basic problem with the standard shared frailty model

$$\lambda(t) = Z_i\, \lambda_0(t) \exp(X_i^{\mathrm{T}} \beta),$$

where Z_i is a random effect that is gamma distributed with mean 1 and variance θ^{-1}, is that we can identify all parameters solely on the basis of univariate survival data. Therefore the variance parameter cannot be interpreted as reflecting only correlation, but it will also reflect lacking fit of the model. Even though it may not be a big problem for multivariate data it is difficult to know and it is clearly a problem with the model. We believe that the two-stage procedure with marginals on a specific form provides a practical solution to this identifiability problem and we may also use non-parametric maximum likelihood estimation techniques for this model.

The authors seem to prefer normal random effects but we see no reason why these should be preferred. It has been shown that different random effects lead to different types of dependence and it will vary from case to case which random-effect distribution leads to the best description of the correlation (Hougaard, 2000).

We have considered the colon cancer data in the case of the shared frailty model that is contained in the class of models that was considered by the authors but with a gamma-distributed frailty for simplicity. The colon study is clearly asymmetric since death will censor cancer whereas the opposite is not true; therefore any correlation should be identified on the basis of how the occurrence of cancer changes the death-rates. The frailty model considered may be written as

$$\lambda_{ic}(t) = Y_{ic}(t) Z_i\, \lambda_{\mathrm{c}}(t) \exp(X_{ic}^{\mathrm{T}} \beta),$$

$$\lambda_{id}(t) = Y_{id}(t) Z_i\, \lambda_{\mathrm{d}}(t) \exp(X_{id}^{\mathrm{T}} \beta)$$

for the ith patient where c is for cancer and d is for death, and Z_i is gamma distributed with mean 1 and variance θ^{-1}. One special feature of the asymmetry is that the at-risk indicator $Y_{ic}(t)$ is 0 when a patient has died but $Y_{id}(t)$ is equal to 1 if the patient is still alive even if the patient has experienced cancer. The intensities with respect to the observed history are

$$\tilde{\lambda}_{ic}(t) = Y_{ic}(t)\, \lambda_{\mathrm{c}}(t) \exp(X_{ic}^{\mathrm{T}} \beta) \frac{1}{1 + \{\Lambda_{ic}(t) + \Lambda_{id}(t)\}\theta^{-1}},$$

$$\tilde{\lambda}_{id}(t) = Y_{id}(t)\, \lambda_{\mathrm{d}}(t) \exp(X_{id}^{\mathrm{T}} \beta) \frac{1 + N_{ic}(t-)\theta^{-1}}{1 + \{\Lambda_{ic}(t) + \Lambda_{id}(t)\}\theta^{-1}}.$$

For the colon data there are no deaths for subjects who did not experience cancer and so all the deaths are for subjects who did experience cancer, thus indicating that a frailty-type model is not well suited to fit these data. Alternatively one may consider intensity models where one directly models the effect of cancer on the death-rate by conditioning on the timing of this event.

Per Kragh Andersen (*University of Copenhagen*)
Professor Zeng and Professor Lin have presented a very impressive paper covering a broad range of models for univariate and multivariate survival data, and repeated events and joint models for longitudinal data and survival data. They could give a unified approach to maximum likelihood estimation in such models, including asymptotic theory, based on results for empirical processes. I would like to address two aspects of the likelihood derivation: one deals with types of observational patterns; the other with types of time-dependent covariates.

Though 'general censoring and truncation patterns' are, indeed, mentioned when presenting the likelihood (5) it seems as if later likelihoods in random-effects models will only be applicable for independent and non-informative (in the sense of Kalbfleisch and Prentice (2002), chapter 6, and Andersen *et al.* (1993), chapter III) *right* censoring. Random left truncation, for example, would require non-identical frailty distributions as the distribution of b_i must be evaluated conditionally on survival beyond the left truncation time. It is not clear whether the theorems as stated in Appendix A cover this situation.

The likelihoods (5) and (8) are only full likelihoods if the time-dependent covariates fulfil certain simplifying assumptions. This could be that they are either deterministic, ancillary or adapted to the history that is generated by the failure time counting process (Kalbfleisch and Prentice (2002), chapter 6, and Andersen *et al.* (1993), chapter III). For *internal* time-dependent covariates likelihoods (5) and (8) are only *partial* likelihoods. In Section 2.3 the resulting likelihoods (9) and (11) may be full likelihoods since a joint model is available for the failure time data and the internal time-dependent covariate. Does the type of likelihood have consequences for the asymptotic results that are derived in Appendix B and for the use of the EM algorithm or do these results only rely on the shape of the likelihood (see equation (12))?

In Section 2.3 covariates X are introduced in the model for response variables Y and it is stated that 'Typically, but not necessarily $X_{ij} = Z(t_{ij})$'. Consider this situation for the simple Cox-type version of equation (2) with a single time-dependent covariate and random effects:

$$\lambda_i(t) = \lambda(t) \exp\{\beta Z_i(t) + \psi_0 b_{i0} + \psi_1 b_{i1} Z_i(t)\}.$$

Consider also the simple linear version of model (10):

$$Y_{ij} = \alpha_0 + \alpha_1 Z_i(t_{ij}) + b_{i0} + b_{i1} Z_i(t_{ij}).$$

To compute the likelihoods (5), (8), (9) and (11) values of $Z_i(t)$ must be observed for all t or, at least, for all observed event times. In contrast, the response variables Y are only observed at certain ('non-informative') measurement times t_{i_j}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. One situation in which it is plausible to observe $Z_i(t)$ continuously but Y only at t_{ij} arises when $Z_i(t)$ is deterministic (see Tsiatis and Davidian (2004)) but if $Z_i(t)$, more generally, is a random process it is not clear whether such observations arise in practice.

V. T. Farewell and B. D. M. Tom (*Medical Research Council Biostatistics Unit, Cambridge*)
We commend the authors on an impressive piece of work. The class of semiparametric regression models proposed provides a comprehensive framework for the identification of relationships between occurrence of events and potential explanatory variables.

As a means for the examination of goodness of fit regression models, and perhaps as the basis of exploratory investigation of relationships along with graphical and tabular procedures, the class is of particular interest. We would like to ask, however, whether, for the reporting of relationships between an outcome of interest and explanatory variables, some caution might be wise. Potential issues relate to interpretability, overfitting and reproducibility.

For the gastric cancer example, the authors' Fig. 1 displays fits from their two heteroscedastic versions of the linear transformation model, and the claim is made that the second version fits better. From an applied perspective, we wonder how important the better fit is in this example. Additionally, we present a figure here that results from fitting a Cox regression model with two explanatory variables, treatment and a log(time) × treatment 'interaction' (Fig. 6). This time-dependent Cox model we infer would be a special case of their class of models. The interaction variable is highly significant and the fit is comparable with those provided in the paper, particularly that of the authors' model (3). We wonder therefore whether this approach to a regression model might provide a suitable representation of the treatment

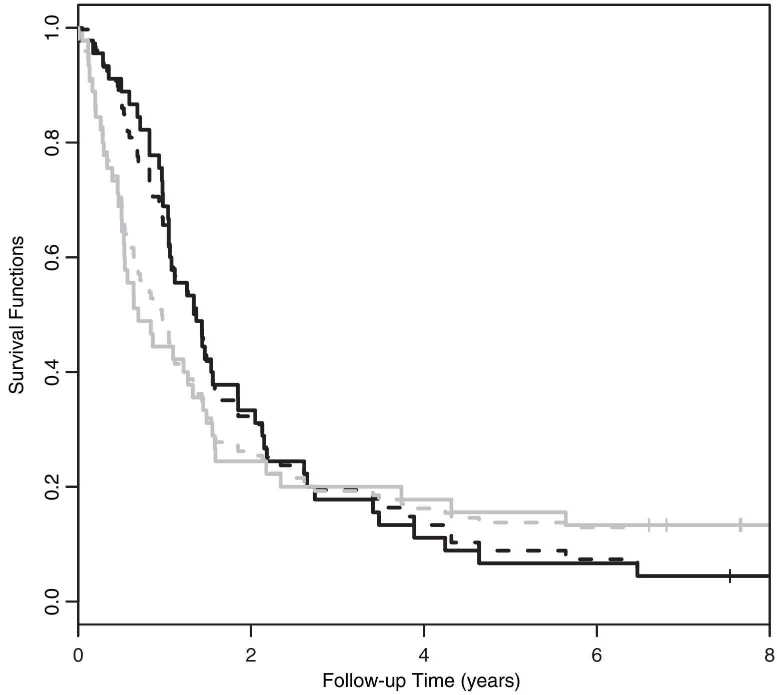

Fig. 6. Kaplan–Meier and time-dependent Cox model estimates of the survival functions for the gastric cancer data set: ———, Kaplan–Meier, chemotherapy; ———, Kaplan–Meier, combined therapy; -------, time-dependent Cox model, chemotherapy; - - - -, time-dependent Cox model, combined therapy

effect for empirical modelling. Would the nature of the treatment effect be conveyed more usefully, in terms of it varying with a simple function of time, than in terms of the β- and γ-parameters of the models proposed?

With respect to overfitting and reproducibility, we wonder whether there is the potential that their models are sensitive to aspects of a particular data set, which might be absent in other studies of the same question. In particular, one might ask whether, for their models (3) and (4), the authors would respond differently to variation in β-estimates than to variation in γ-estimates across studies.

In addition, the simulation studies in the paper are restricted to data that were drawn from their proposed class of models. The behaviour of this class in other situations (e.g. data drawn from Aalen's additive hazards model) might be of interest.

Again, we congratulate the authors and ask these questions to understand better how the approach that they have developed can be incorporated into the current body of techniques that are used for comparable problems.

D. R. Cox (*Nuffield College, Oxford*)

It is a pleasure to have the chance of congratulating the authors on an interesting and valuable paper. It raises many points: some detailed and some general.

That fine Mexican statistician, the late Francisco Aranda-Ordaz, was the first to study estimated transformations in this context, although in a way that was different from that used in the present paper (Aranda-Ordaz, 1983). The simplest way to test for and to represent non-proportionality of hazards is often through a manufactured time-dependent variable (Cox, 1972; Grambsch and Therneau, 1994). Table 2 illustrates that, although regression coefficients in different models have different interpretations, ratios of regression coefficients are relatively stable. There is a simple invariance-based qualitative explanation of this; for a theoretical treatment in the present context, see Solomon (1984) and Struthers and Kalbfleisch (1986).

Broader aspects are why the hazard and are there special reasons for proportionality? Failure is a stochastic process, Markovian if properly described, and stochastic processes are usually best specified by their transition probabilities, in this case the hazard, which is the complete intensity function of a point process. Note, though, that, although the accelerated life model, which is very natural in some physical contexts with a single failure mode, can be specified via the hazard, that is not the most direct definition. Proportional odds and cumulative hazards seem much less natural as a base for initial specification.

The contrast between proportional and additive measures pervades epidemiology. Proportionality evades positivity constraints, specifies effects essentially in dimensionless form and possibly gives more stability in estimated effects. Yet quite clearly there are no all-persuasive arguments for proportionality.

More broadly still, the paper illustrates a wide-ranging tension in statistical development. If models are compact essentially descriptive representations of patterns of variability, the move to ever more general families of models is a very welcome and fruitful one. If the objective is in part to probe the data-generating process and to provide simple more incisive interpretations, more specificity in model specification may often be preferable (Cox, 1972) and this may or may not lie naturally within some prescribed general setting.

Ørnulf Borgan (*University of Oslo*)

I congratulate the authors on an impressive paper that will have an influence on the further development of survival and event history analysis.

In Section 1 the authors praise the 'ingenious partial likelihood principle' and the 'elegant counting process martingale theory'. However, a main message is that inference for survival and event history models should be based on non-parametric maximum likelihood estimation and that empirical processes are the preferred mathematical tool. I shall advocate a more pragmatic attitude: we should adopt the inference methods and mathematical tools that are most convenient, seen both from a theoretical and a practical perspective. I shall use my experience with cohort sampling methods to underpin my point of view.

There are two classes of cohort sampling methods: the nested case–control and case–cohort designs. These designs are only briefly mentioned in the paper but, as they are likely to gain importance, methods for survival and event history data should be able to address the methodological problems of cohort sampling methods.

In a nested case–control study, a few controls are sampled from those at risk at the failure times. A joint model for the occurrences of failures and the sampling of controls may be formulated and studied by using counting processes and martingales (Borgan *et al.*, 1995). By allowing the sampling probabilities to depend on covariates for *all* individuals at risk (thus leaving an independent and identically distributed data set-up), the framework makes it possible to tailor the control sampling to the specific design and analysis needs of a study (e.g. Langholz (2007)). Inference may be based on a Cox-type partial likelihood and easily performed by using standard software. For *simple* nested case–control designs, alternative inference methods have been suggested (e.g. Scheike and Juul (2004) and Samuelsen *et al.* (2007)). However, in my opinion, the modest gains in efficiency that are obtained by these methods (for simple nested case–control designs) do not outweigh the practical complications in fitting the models and the loss of flexibility in tailoring the control sampling to the study needs.

In a case–cohort study, a subcohort is sampled at the outset of the study by simple or stratified sampling, and the at-risk individuals in the subcohort are used as controls at all failure times. For case–cohort designs the partial likelihood degenerates, and estimation is usually based on a pseudolikelihood. Counting processes and martingales are of no help in studying the properties of the estimators, which must be done by using empirical processes for finite population sampling (e.g. Breslow and Wellner (2007)).

P. M. Philipson and W. K. Ho (*Newcastle University*)

The paper builds on oft-used models in an interesting and widely applicable manner and, as such, the authors are to be commended. Our comments concern the survival models of Section 2.1.

We feel that there is scope for a link between the models that are routinely used at present and those that are proposed in this paper. Given the overwhelming popularity of the Cox proportional hazards model, a useful intermediate step would be to ascertain whether there is sufficient evidence to warrant the use of transformation models, or the extension to the crossing hazards model. Score tests, requiring only the trivial fitting of a proportional hazards model, could be used for such a purpose.

Consider the class of Box–Cox transformations. If we assume that only time invariant covariates are present and that β and Λ_0 are known, then, appealing to martingale theory, we obtain

$$U_{\rho 0}=\sum_i \left[\int_0^\tau \log\{1+\Lambda_i(u)\}\,\mathrm{d}M_i(u)\right]$$

as the score for the transformation parameter ρ under the null hypothesis (i.e. $\rho=\rho_0=1$), where M_i is the usual counting process martingale and τ is the duration of study. The predictable variation under the null hypothesis is

$$V_{\rho 0}=\sum_i \left[\int_0^{\tau} \log^2\{1+\Lambda_i(u)\}\,\mathrm{d}\Lambda_i(u)\right].$$

Focusing on the model given by equation (4) of the paper, ignoring any transformations for now, allows investigation of the crossing hazards case. In an analogous fashion, the score for γ (here assumed for convenience to be scalar), under the null hypothesis, can be expressed as

$$U_{\gamma 0}=\sum_i \left(\int_0^{\tau} \tilde{Z}_i[1+\log\{1+\Lambda_i(u)\}]\,\mathrm{d}M_i(u)\right)$$

with associated variance

$$V_{\gamma 0}=\sum_i \left(\int_0^{\tau} \tilde{Z}_i^2[1+\log\{1+\Lambda_i(u)\}]^2\,\mathrm{d}\Lambda_i(u)\right).$$

This fledgling idea is at a very embryonic stage; clearly adjustments will need to be made to accommodate estimation of β and Λ_0. It is hoped that fully developed tests would dovetail with the novel models that have been put forward in this paper and provide clarity for the statistician.

The two cases of transformation and crossing hazards have been considered separately here. The similarity of the above expressions for ρ and γ leads us to ponder the effect of fitting models when both cases are present. Are the parameters suitably disentangled so that estimation remains robust? Do the authors have any insight to offer, or any reflections on the fitting of such models?

John A. Nelder (*Imperial College London*)
Those of us who have been working on *h*-likelihood methods (Lee *et al.*, 2006) are naturally disappointed to see no reference to them in this paper. Professor Lee and Professor Ha will give detailed comments on the use of *h*-likelihood for fitting the model class that is described in the paper. I shall just say that more general frailty models, allowing structured dispersion, can be expressed in the form of hierarchical generalized linear models after a suitable arrangement of the data matrix (Noh *et al.*, 2006). Advantages of this approach in fitting the models are the following.

(a) Quadrature is not required.
(b) The EM algorithm, that lumbering giant of an algorithm, is not required.
(c) The bias in the frailty coefficient that is caused by the large number of parameters is no longer a problem.
(d) Standard software (in Genstat) is available to fit these models efficiently.

I commend this approach to the authors.

J. L. Hutton (*University of Warwick, Coventry*)
I thank the authors for a thought-provoking paper. I make three simple remarks as a medical statistician.

Accelerated life models are useful in medical applications. In my work on cerebral palsy and epilespy, accelerated life models have been more useful than proportional hazard models (Hutton and Pharoah, 2002; Hutton and Monaghan, 2002; Cowling *et al.*, 2006). Professor Henderson mentioned that hazards often converge over time and suggested that proportional hazard models needed to incorporate time-dependent elements to allow for this. Accelerated life models with a log–logistic or gamma base-line have converging hazard ratios. With a log–logistic base-line, the hazard ratio converges to 1.

Hazard functions should not cross. My response to the data on gastric cancer patients is to ask whether an oncologist or pharmacologist could suggest covariates which would distinguish those patients who die early on from those with a better prognosis. Accelerated life models with Pareto distributions were effective in allowing us to understand the responses to antiepileptic drugs (Cowling *et al.*, 2007).

The following contributions were received in writing after the meeting.

Peter J. Bickel (*University of California, Berkeley*)
I enjoyed the authors' presentation of a general toolbox of models for censored survival data. Their paper raised an old philosophical issue for me.

The philosophical point has to do with the failure to account for the variability of the transformation parameter estimates in the confidence bands for the treatment effect parameter that they represent. This was the subject of Bickel and Doksum (1981) and a subsequent discussion of a paper of Hinkley and Runger (1984). As I agreed, stating confidence bounds on an effect on an unknown scale is problematic, but so is underestimating variability. The simple solution is to announce only simultaneous confidence limits for the transformation parameters and the effects, permitting the reader to interpret the transformation effect on their selected scale.

Here are a few more comments.

(a) The authors point out that non-parametric maximum likelihood estimators are intractable computationally, leading to *ad hoc* estimates. The *ad hoc* estimates may be the most reasonable starting-point for their optimization procedure. If it involves Newton–Raphson iteration, a single step from an $o(n^{1/4})$ consistent estimate will suffice for efficiency as remarked by LeCam and extended in Bickel (1975).

(b) I want to stress a point that was alluded to by the authors in their data analysis. If, as expected, the model is a crude approximation to the mechanism producing the data, the interpretability of quantities that are estimated is important. In the authors' models the form of G and ε must be predetermined. What they estimate are distributions that are closest to the truth in Kullback–Leibler distance—but what this means for parameters of interest is not necessarily clear.

(c) The foregoing suggests that making G and even the distribution of ε non-parametric is worthwhile if only to see what effect this has on the parameters of interest. As the authors point out, not specifying the distribution of ε, in the models of Section 2.1, makes the parameter of interest β unidentifiable—but only in a weak sense. We can still identify the components of β up to a common scale factor, so relative magnitudes of effects can be measured. Econometricians have studied this problem in simpler contexts as discussed in Horowitz (1998).

(d) Finally, here is a note of caution. As the complexity of the model increases, so does the number of parameters explicit or implicit as in the estimated H. I believe that one needs to think about imposing sparsity on one's models.

N. E. Breslow (*University of Washington, Seattle*)

I congratulate Zeng and Lin for their construction of a general model that nicely brings together much previous work, their development of asymptotic theory that justifies inference using profile likelihood and their description of innovative computational approaches. They provide little guidance, however, regarding parameter interpretation. One important benefit of semiparametric models is that quantities that are of key scientific interest may be summarized parametrically whereas nuisance factors are treated non-parametrically. Thus the Cox model focuses attention on the well-understood hazard ratio. Marginal mean and marginally specified hierarchical models (Heagerty and Zeger, 2000) express treatment effects in terms of population averages, possibly within subgroups defined by covariates, whereas marginal structural models (Robins *et al.*, 2000) express population level effects that would be observed if treated and control groups had the same covariate distribution. Parameters in hierarchical transformation models, by contrast, must be interpreted conditionally and may be highly sensitive to distributional assumptions. How to interpret the fixed covariate coefficients in the random-effects transformation model for longitudinal data, where the link is left unspecified, seems particularly deserving of comment.

When faced with crossing hazards, as in Fig. 1, we might add a time-dependent covariate to the Cox model and express the log-hazard-ratio parametrically as $\beta_0 + \beta_1\, g(t)$ where $g(\cdot)$ is a specified, increasing function such as $g(t) = \log(t/t_0)$ with t_0 a modal time. Then β_0 expresses the log-hazard-ratio at t_0 and β_1 its rate of change with a known function of time. With the heteroscedastic transformation model, by contrast, $g(t) = H(t) = \log\{\Lambda(t)\}$ and the interpretation is less clear.

For missing data problems, including two-phase studies where data are missing by design, Zeng and Lin note that Horvitz–Thompson or inverse probability of sampling weighted estimators may be more robust than non-parametric maximum likelihood estimators in the face of model misspecification. Indeed, survey statisticians advocate their use on grounds that they consistently estimate the parameters obtained when fitting a possibly misspecified model to the source population, which the non-parametric maximum likelihood estimator may fail to do. Breslow and Wellner (2007) provide theory for Horvitz–Thompson estimation of both Euclidean and infinite dimensional parameters in semiparametric models fitted to data from two-phase stratified samples. Efficiency may be enhanced through adjustment of the sampling weights by using the survey techniques of post-stratification and calibration, or by their estimation. I concur wholeheartedly that further studies are needed to assess the relative merits of Horvitz–Thompson and non-parametric maximum likelihood methods for complex sampling designs.

Jack Cuzick (*Wolfson Institute of Preventive Medicine, London*)
We have recently extended our study of non-compliance in randomized trials from the binary outcome case (Cuzick *et al.*, 1997) to a proportional hazard set-up (Cuzick *et al.*, 2007). This greatly complicates estimation, and we have developed a non-iterative *ad hoc* estimator, as well as estimators that are based on a partial likelihood and the full semiparametric likelihood. The complications arise because non-compliance is modelled by a covariate (insistor, refuser or ambivalent) which is incompletely observed, leading to a latent class model. In particular insistors cannot be distinguished from patients who are willing to accept either treatment in the active treatment arm, whereas refusers and ambivalent patients are indistinguishable in the control arm. In simulations we have found that the non-parametric maximum likelihood estimator that is computed by using the BFGS quasi-Newton algorithm outperforms other estimators in a wide range of conditions, but an asymptotic theory has eluded us. The likelihood is of their general form (12) but not the specific form (11), and validation of the conditions of their general theorem in this case is still formidable, as is computation of the asymptotic variances. Nevertheless the paper gives a very valuable foundation for studying a wide range of semiparametric problems and will no doubt become a standard reference for further research.

Paddy Farrington (*The Open University, Milton Keynes*) **and Mounia Hocine and Thierry Moreau** (*Institut National de la Santé et de la Recherche Médicale, Paris*)
The authors have achieved an impressive unification and extension of several classes of semiparametric models for event history and repeated measures data. Particularly useful is the general asymptotic theory underpinning these models.

Does the modelling framework for recurrent events encompass models for different timescales, including both time from last event and calendar time? We were mystified by the comment regarding clinical trials at the beginning of Section 2.2. For example, Duchateau *et al.* (2003) proposed a model involving both frailties and time-varying effects in the form of gap time-dependent hazards, which they applied to clinical trial data.

General and complex models induce problems of model identification and interpretation. For example, crossing hazards may be attributable to one of several contrasting effects. These include time varying exposures, selection effects and the functional form of the dependence of the hazard on fixed covariates. All three are available in the present models: to what extent are they identifiable from data?

Furthermore, interpreting the model parameters is far from easy, as illustrated by several of the examples in the paper. In the gastrointestinal tumour example, the estimated parameter value for the treatment effect for the well fitting model (4) is $\beta = 3.028$ (0.262). Yet to make sense of this requires us to look at the Kaplan–Meier survival curves: knowledge of β provides little further enlightenment. A similar point can be made about the colon cancer example: the parameter estimates for the selected model provide little clue about how effective the treatment really is.

It would be useful in particular if the authors could clarify under what parameter combinations the hazards do not cross, and whether a test for non-crossing hazards could be derived for model (4) which could lead to simplifications. For example, in model (3) with $G(x) = x$ and fixed covariates, the null hypothesis of no crossing is simply $\gamma = 0$, and the corresponding score test is readily obtained (Quantin *et al.*, 1996).

An alternative to building ever more complex models is perhaps to focus on the questions of primary interest, while eliminating nuisance parameters by conditioning. One such approach, admittedly for much simpler data structures than those considered here, is provided by the semiparametric case series model (Farrington and Whitaker, 2006). This employs a conditioning argument to eliminate the multiplicative effects of frailties and non-varying covariates, thus focusing the analysis on time varying exposures of interest.

Jason P. Fine (*University of Wisconsin, Madison*)
Theory for non-parametric maximum likelihood estimation has percolated over the past decade, stimulated by the seminal work of Murphy (l994, 1995). The current paper presents potentially useful albeit somewhat straightforward extensions, with the modelling ideas and theoretical developments following closely earlier contributions. The argument that such methodology should play a wider role in statistical practice is intriguing. Unfortunately, the rationale for widespread adoption in applications is less

convincing than that for the underlying mathematics. There is an inattention to key applied issues and the relevance and practical utility of the framework in the examples.

In the cancer illustration in Section 5.1, the treatment effect clearly violates proportional hazards and fitting models (3) and (4) yields very different results from a naïve proportional hazards analysis. Viewing Fig. 1, the study investigators would be interested in understanding how the treatment effect changes over time in the population. This is obscured by the heteroscedastic transformation model, whose interpretation is rather mathematical and difficult for non-statisticians. The proportional hazards model easily accomodates time-dependent effects. Either time-dependent covariates involving interactions of treatment and time (Therneau and Grambsch, 2000) or time-dependent coefficients (Martinussen and Scheike, 2006) may be employed.

In Section 5.2, the joint model for the recurrence time X and the death time Y is questionable. The resulting analysis of the marginal distribution of recurrence corresponds to the setting where death before recurrence does not occur, which is generally of secondary interest to patients and physicians. Typically, competing risk end points, like cause-specific hazard and cumulative incidence (Kalbfleisch and Prentice, 2002), are reported in oncology journals. Their interpretation corresponds to the current reality where death may occur before recurrence, which is of greater interest. If the distribution of residual life post recurrence is of interest, then the distribution of $Y-X$ can be directly modelled conditionally on X and other covariates using proportional hazards models, which would be the default in practice. This simple analysis can be implemented using standard software and the interpretation of the effect of X on $Y-X$ is much more transparent than that in the joint model.

For multivariate data, random-effects models may be useful for assessing failure time correlations. The gamma distribution is attractive, because of its relationship to the cross-hazard ratio (Oakes, 1989). The model can be generalized to permit time varying and asymmetric associations. The cross-hazard interpretation is opaque for the normal distribution (Hougaard, 2000). When correlation is a nuisance, random-effects models seem less attractive, as their misspecification may bias other parameter estimators. Moreover, incorporating random effects in conditional proportional hazards models generally gives marginal non-proportional hazards models, whose interpretation may be problematic. Marginal proportional hazards models avoid such limitations. These models can be coupled with copulas, e.g. multivariate normal. Presumably, non-parametric maximum likelihood estimation inferences are efficient, similarly to random-effects models.

Il Do Ha (*Daegu Hanny University, Daegu*)

For the maximum likelihood (ML) estimation the authors use the EM algorithm and the discrete non-parametric Breslow estimator, which results in biased estimators (Rondeau *et al.*, 2003). Overall, the authors consider bivariate survival data, where there is less of a problem than with univariate data (Barker and Henderson, 2005). We now demonstrate how the h-likelihood approach overcomes this problem. For simplicity of argument, we consider the semiparametric frailty models (1) with clustered failure time data. We assume that u_i has a gamma distribution with $E(u_i)=1$ and $\text{var}(u_i)=\alpha$ to allow an explicit marginal log-likelihood m.

Let m^* be the profile marginal likelihood (Nielsen *et al.*, 1992; Murphy and van der Vaart, 2000) after eliminating the nuisance parameter λ_0, defined by

$$m^* = m|_{\lambda_0=\tilde{\lambda}_0},$$

where $m=\log\{\int\exp(h)\,\mathrm{d}v\}$ is the marginal likelihood, h is the h-likelihood (Lee and Nelder, 1996; Ha *et al.*, 2001) and $\tilde{\lambda}_0$ is the discrete Breslow estimator, obtained from $\partial m/\partial\lambda_0=0$. In fact, the maximization of m^* gives the ML estimators by using the EM algorithm (Andersen *et al.*, 1997). The resulting ML estimators have downward biases, particularly for the frailty parameter α.

On the basis of 200 replications of simulated data we investigate the performances of three profile likelihood methods (m^*, $p_w(m)$ and $p_v^s(h^*)$). For the gamma frailty we use the second-order Laplace approximation $p_v^s(h^*)$ (Lee and Nelder, 2001). Given the frailty parameter α, we use profile likelihoods m^* and h^*, which provide the same estimates for β (Ha *et al.*, 2001). However, they give different estimators for α because the estimates of α are obtained by maximizing the three adjusted profile likelihoods. Under no censoring we generate data by assuming the exponential base-line hazard $\lambda_0(t)=1$, one standard normal covariate with $\beta=1$, and $\alpha=1$. We consider both univariate and bivariate sample cases: $N=\Sigma_{i=1}^n n_i=(100,200)$ with $(n,n_i)=(100,1),(100,2),(200,1)$. Note here that we choose fairly extreme cases, with no censoring and small sample size, because these situations yielded the most biased estimates of $\hat{\alpha}$ in the simulation studies by Nielsen *et al.* (1992) and Barker and Henderson (2005). The results are summarized in Table 5.

Table 5. Simulation results for the estimators $\hat{\alpha}$ and $\hat{\beta}$ under marginal and h-likelihoods in semiparametric gamma frailty models†

n	n_i	*Method*	*Results for* $\hat{\alpha}$			*Results for* $\hat{\beta}$		
			Mean	*Standard deviation*	*Mean-squared error*	*Mean*	*Standard deviation*	*Mean-squared error*
100	1	m^*	0.42	0.363	0.469	0.80	0.228	0.094
		$p_w(m)$	0.87	0.618	0.398	0.97	0.297	0.088
		$p_v^{\mathrm{s}}(h^*)$	0.89	0.617	0.393	0.96	0.296	0.089
100	2	m^*	0.90	0.240	0.067	0.98	0.281	0.079
		$p_w(m)$	0.99	0.258	0.066	1.00	0.286	0.081
		$p_v^{\mathrm{s}}(h^*)$	0.99	0.258	0.066	1.00	0.286	0.081
200	1	m^*	0.63	0.301	0.231	0.87	0.188	0.054
		$p_w(m)$	0.98	0.437	0.191	0.99	0.221	0.050
		$p_v^{\mathrm{s}}(h^*)$	0.99	0.439	0.192	1.00	0.231	0.053

†The simulation is conducted with 200 replications at true gamma frailty variance $\alpha = 1$ and regression parameter $\beta = 1$ (no censoring).

As expected, m^* gives severe downward biases in all cases considered, especially with $n_i = 1$. Moreover, the underestimation of α leads to that of β. Table 5 also demonstrates that the two adjusted profile likelihoods $p_w(m)$ and $p_v^{\mathrm{s}}(h^*)$ reduce such biases substantially, giving almost the same results.

Here we have considered the gamma frailty model to have an explicit form for m and $p_w(m)$. However, this is not so in general, for example, for models with log-normal frailty, or with nested and/or serially correlated frailty. Thus, the adjusted profile likelihoods that are based on h-likelihood are useful for general frailty models. We believe that the h-likelihood approach gives more flexible ML inferences than the EM approach.

Joel L. Horowitz (*Northwestern University, Evanston*)
I congratulate Professor Zeng and Professor Lin on their interesting paper. It presents a class of flexible semiparametric models for censored survival and longitudinal data. The models accommodate a wide variety of distributions of random effects or frailty. In particular, the standard assumption of gamma-distributed random effects is removed. It is useful to ask whether the assumptions about frailty can be relaxed further by making the frailty distribution non-parametric.

There has been much interest in this question in econometrics over the past two decades. Heckman and Singer (1984a) showed that the parameter estimates from a Weibull hazard model are very sensitive to the choice of frailty distribution. They established consistency of a non-parametric maximum likelihood estimator of this distribution. Elbers and Ridder (1982), Heckman and Singer (1984b) and Ridder (1990) gave conditions for identification of proportional hazard and generalized accelerated failure time models with non-parametric frailty. Honoré (1990) developed an estimator of the shape parameter of a Weibull hazard model with frailty and gave conditions under which it is asymptotically normal with a rate of convergence in probability that is arbitrarily close to $n^{-1/3}$. Horowitz (1999) showed how to estimate a proportional hazard model in which the base-line hazard function and frailty distribution are both non-parametric. Horowitz's estimator, like Honoré's, has a slower than $n^{-1/2}$ rate of convergence. This happens because identification is through the behaviour of the hazard function in an arbitrarily small neighbourhood of 0. However, Ridder and Woutersen (2003) showed that $n^{-1/2}$-convergence is possible if the base-line hazard function is bounded away from 0 and ∞ in a neighbourhood of 0. An $n^{-1/2}$ rate of convergence is also possible if we have longitudinal data (Horowitz and Lee, 2004). Indeed, this is possible with longitudinal data even if the frailty variable is correlated with the covariates.

Non-parametric estimation of the frailty distribution with cross-sectional data is a deconvolution problem, so the rates of convergence in probability are quite slow in general. None-the-less, it appears possible to obtain useful estimates with samples of practical sizes (Horowitz, 1999). In summary, estimation results in hazard models can be very sensitive to misspecification of the frailty distribution. Non-parametric treatment of the frailty distribution is possible in simple models such as proportional hazards models. It is worth investigating whether more complicated models such as those of Zeng and Lin are also sensitive to misspecification of the frailty distribution and whether non-parametric estimation of this distribution is possible.

John D. Kalbfleisch (*University of Michigan, Ann Arbor, and National University of Singapore*) **and Jinfeng Xu** (*National University of Singapore*)
We congratulate Professor Zeng and Professor Lin on an interesting and far reaching paper with many facets and intricacies. Our allotted space is very short, so we confine our comments to three points.

There is clearly value in developing methods for and investigating uses of alternative models but, like many others, this paper begins by setting up proportional hazards as a straw man. There is no recognition that the covariates in the model can incorporate interactions with functions of time and so model, quite parsimoniously, various non-proportional aspects of a problem. For this reason, Kalbfleisch and Prentice (2002) suggest use of the term relative risk or Cox model instead of the proportional hazards misnomer. The proposal of incorporating heterogeneous errors in a linear transformation model to account for possible non-proportional hazards leads to incorporation of coefficients that are difficult to interpret, and more difficult, we suggest, than the interpretation of a time varying term in a relative risk model.

How in general do we interpret parameters in a linear transformation model? There are simple interpretations for extreme value or logistic error. In the more general case with estimated G and arbitrary H, however, it is not clear what β is measuring in model (1) or its extensions. Table 1 illustrates the point. Here the parameters in the relative risk model have simple interpretations as log-relative-risks and similar interpretations apply in the proportional odds model. The paper notes that the 'interpretation of treatment effects ... depends on which model is used' but provides no guidance on interpreting the parameters under the suggested analysis. We seem to be left with a test for no treatment effect but without the benefit of interpretable parameters.

Finally, Fig. 2 comprises four separate sheets wherein (0,0), (0,40), (40,0) and (40,40) all correspond to proportional odds and (20,20) corresponds to proportional hazards. It is not clear that (20,20) corresponds to a single ordinate in all four sheets as it should; nor is it apparent that all representations of proportional odds (e.g. (0,40) and (0,0)) have the same ordinate, and we wonder whether the normalizing constant is the same in all sheets. A contour plot would perhaps have been more informative. The reason for the arbitrary cut-off at $\rho = 1$ also is unclear and seems that it may affect the model selected. Some investigation of simple time-dependent relative risks in this example would be interesting.

Michael R. Kosorok (*University of North Carolina, Chapel Hill*)
I congratulate Zeng and Lin on an excellent contribution to statistical modelling for right-censored data. The authors make a very strong case for the practical use of efficient, maximum-likelihood-based estimation for semiparametric models. Moreover, the heteroscedastic linear transformation model proposed and, especially, the random-effects linear transformation model are scientifically interesting and very appealing new models.

Nevertheless, there are a few points to be made. To begin with, several important references should be added to the part of the introduction that reviews transformation models. Slud and Vonta (2004) generalized the work of Scharfstein *et al.* (1998) to more general choices of the G-function than is given in expression (2) of the paper under discussion. Kosorok *et al.* (2004) further generalized to allow G to be parameterized with unknown parameter values. Thus the future topic that Zeng and Lin propose in the second paragraph of Section 7 has already been partly accomplished in Kosorok *et al.* (2004).

On a more favourable note, the ensemble of numerical tools that were developed by Zeng and Lin is a key contribution that makes the methods proposed usable in practice. However, additional gains in computational efficiency are possible when both the finite and the infinite dimensional parameters are jointly efficiently estimated, as has been verified, for example, by Kosorok *et al.* (2004) for transformation models with right-censored data. Incidentally, this joint efficiency holds for most of the models in Zeng and Lin's paper, even though they neglected to point this out. This gain in computational efficiency is achievable through careful utilization of the profile likelihood structure for both the finite dimensional parameter via the profile sampler (Lee *et al.*, 2005) and for all parameters jointly via the piggyback bootstrap (Dixon *et al.*, 2005). The computational savings of these methods have been verified rigorously and can be dramatic. Combining the profile sampler and piggyback bootstrap with the numerical innovations of Zeng and Lin should lead to further dramatic improvements.

In Section 7, the authors mention robustness under model misspecification and extending transformation models to interval-censored data as important future topics. Some initial work on robustness of transformation models was given in Kosorok *et al.* (2004), who showed that the direction of the regression effects can be accurately estimated even when the G-function is misspecified. Likelihood inference under interval censoring for transformation models with partly linear regression effects was developed and verified theoretically in Ma and Kosorok (2005). A key challenge here is the presence of two non-root-n consistent estimators.

In general, very little work has been done for likelihood-based semiparametric inference involving parameters that are not root n estimable. A very careful analysis of the entropy of these models is usually required, making this area of endeavour one of the most intellectually demanding in all of statistics. Adding to this the other open topics that were mentioned by Zeng and Lin, it is clear that many challenging questions in semiparametric inference remain.

Youngjo Lee (*Seoul National University, Seoul*)
I congratulate the authors on unifying maximum likelihood estimation for the analysis of multivariate survival data, based on the EM algorithm. However, the EM method is slow and may not be easily applicable to complicated situations. For simplicity of argument, consider the semiparametric frailty models (6) with clustered failure time data, giving conditional hazard

$$\lambda(t|Z_{il};u_i) = \lambda_0(t)\ \exp(\beta^{\mathrm{T}} Z_{il})u_i, \tag{16}$$

where $\lambda_0(\cdot)$ is an unspecified base-line hazard function and the u_i follow some distribution. For inferences, Lee and Nelder (1996) proposed to use the h-likelihood, which is defined by

$$h = \log\{f_\theta(y|u)\} + \log\{f_\theta(v)\}$$

where $f_\theta(y|u)$ and $f_\theta(v)$ are probability density functions for $y|u$ and $v = \log(u)$ respectively. For inferences about the fixed parameters θ, the marginal (log-)likelihood has been proposed, using

$$m = \log\{f_\theta(y)\} = \log\left\{\int \exp(h)\,\mathrm{d}v\right\}.$$

However, in general the required integration is intractable. Thus, Lee and Nelder (2001) considered a function class that they called adjusted profile likelihoods $p_\tau(l)$; these eliminate the nuisance parameter τ from a likelihood l, defined by

$$p_\tau(l) = (l - \tfrac{1}{2}\log[\det\{D(l,\tau)/2\pi\}])|_{\tau=\hat{\tau}}$$

where $D(l,\tau) = -\partial^2 l/\partial\tau^2$ and $\hat{\tau}$ solves $\partial l/\partial\tau = 0$. The adjusted profile function $p_v(h)$ eliminates the random parameters v by the Laplace approximation to integration (Lee *et al.*, 2006) and $p_\beta(m)$ eliminates the fixed parameters β by conditioning on $\hat{\beta}$ (Cox and Reid, 1987).

For frailty models, Ha and Lee (2005) proposed to use an adjusted profile h-likelihood $p_v(h^*)$, which is defined by

$$p_v(h^*) = (h^* - \tfrac{1}{2}\log[\det\{D(h^*,v)/2\pi\}])|_{v=\hat{v}},$$

where $h^* = h|_{\lambda_0=\hat{\lambda}_0}$ is a profile h-likelihood with a solution $\hat{\lambda}_0$ from $\partial h/\partial\lambda_0 = 0$, $D(h^*,v) = -\partial^2 h^*/\partial v^2$ and $\hat{v}$ solves $\partial h^*/\partial v = 0$. Ha and Lee (2007) showed that

$$p_v(h^*) \simeq p_w(m),$$

where $w = \log(\lambda_0)$. These adjusted profile likelihoods give practically satisfactory estimators (Ha and Lee, 2005, 2007). Instead of using the E-step the h-likelihood method directly maximizes various adjusted profile likelihoods: for its advantages see Lee *et al.* (2006).

Yi Li (*Harvard School of Public Health and Dana–Farber Cancer Institute, Boston*)
Zeng and Lin are to be congratulated for a wonderful work on non-parametric maximum likelihood estimation for semiparametric frailty regression models. In this comment, I concentrate on the interpretation of frailties.

In the framework of random-effects models, the frailties have been introduced to model the clustering effect and will be useful for prediction as illustrated in Section 5.2. However, they are meant to model the within-cluster dependence as the variance components of the frailties typically gauge the magnitude of such dependence (Diggle *et al.*, 1994). This, however, was not elucidated in this paper. This note bridges the frailty parameters with within-cluster dependence measures and highlights a challenge in interpreting these parameters. To convey the idea, consider a (much) simplified version of model (7) for bivariate failure times (T_1, T_2) with no covariates, namely

$$\Lambda(t|b) = G\left\{ \int_0^t \exp(b)\, \mathrm{d}\Lambda(s) \right\} \tag{17}$$

where $b \sim f(\cdot;\gamma)$. Our goal is to link the variance component γ to a 'model-free' and standardized dependence measure that is commonly used for bivariate survival. One such device is Kendall's coefficient of concordance (Kendall's τ), which can be evaluated by

$$\tau = 4\int_0^\infty \int_0^\infty p(t_1,t_2)\, S(t_1,t_2)\, \mathrm{d}t_1\, \mathrm{d}t_2 - 1$$

where $p(t_1,t_2)$ and $S(t_1,t_2)$ are the joint bivariate density and survival functions respectively (see, for example, Hougaard (2000)). It follows that the joint survival under model (17) is

$$S(t_1,t_2) = \int \exp[-G\{\exp(b)\,\Lambda(t_1)\} - G\{\exp(b)\,\Lambda(t_2)\}]\, f(b;\gamma)\, \mathrm{d}b,$$

and $p(t_1,t_2)$ can be conveniently evaluated by $p(t_1,t_2) = \partial^2 S(t_1,t_2)/\partial t_1\, \partial t_2$. Therefore, γ can be viewed to characterize the bivariate dependence through

$$\tau = 4\int_0^\infty \int_0^\infty \left(\int G'\{\exp(b)t_1\}\, G'\{\exp(b)t_2\} \exp(2b) \exp[-G\{\exp(b)t_1\} - G\{\exp(b)t_2\}]\, f(b;\gamma)\, \mathrm{d}b \right) \times \left(\int \exp[-G\{\exp(b)t_1\} - G\{\exp(b)t_2\}]\, f(b;\gamma)\, \mathrm{d}b \right) \mathrm{d}t_1\, \mathrm{d}t_2 - 1.$$

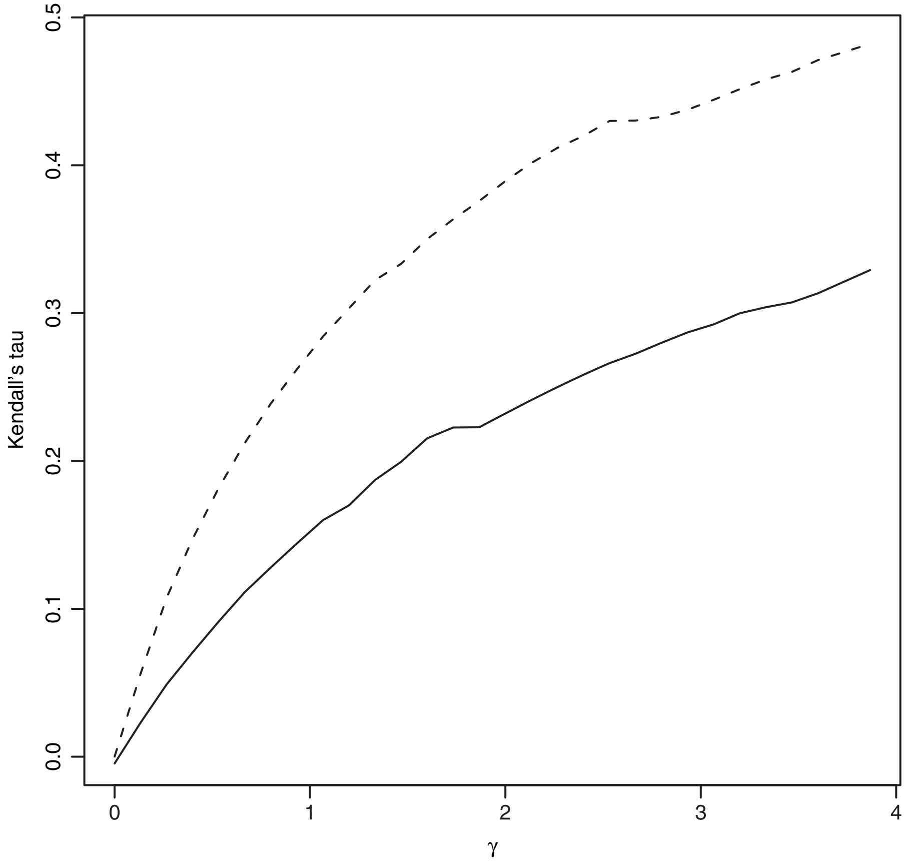

Fig. 7. Kendall's τ *versus* γ for the proportional odds (———) and proportional hazards (– – –) models

It is worth noting that τ does not depend on the base-line function $\Lambda(t)$ in model (17) and its efficient estimate can be obtained by replacing γ with its maximum likelihood estimator $\hat{\gamma}$, whose variance estimate will be immediately available via the delta method. In a similar fashion, the relationship of variance component γ with the other global dependence measures, e.g. the Spearman correlation, integrated hazard ratio and median concordance, and the local dependence measure, i.e. the local cross-ratio, can also be established.

However, a serious challenge of interpreting γ as a dependence measure lies in its dependence on the link function G in model (17). This can be illustrated by Fig. 7, which depicts Kendall's τ against various γ when $b \sim N(0, \gamma)$, under the proportional hazards model (with $G(x) = x$) and the proportional odds model (with $G(x) = \log(1+x)$). For example, $\gamma = 1.8$ corresponds to Kendall's τ of 0.40 under the proportional hazards model, which is almost twice as much as that of 0.21 under the proportional odds model, begging the *cliché* question of 'how large is large?' when viewing the variance component as a measurement for dependence under various transformation models. I would welcome the authors' comments on this issue.

N. T. Longford (*SNTL, Reading, and Universitat Pompeu Fabra, Barcelona*)

I am perplexed by the double negative ('no reason for not using maximum likelihood estimation') in the penultimate sentence of the summary, which I regard as vacuous. The authors do not mention any alternative, and I think that there is no credible alternative to maximum likelihood estimation. The difficulty is not in maximizing a likelihood, such as expression (12), but in specifying one appropriately through the details of the adopted model, balancing the requirements of validity and parsimony. This highlights the need for model selection, or dealing with model uncertainty, and for taking account of the model selection process in the analysis. Information criteria, such as Akaike's information criterion and the Bayes information criterion, are standard but exceedingly poor solutions if the model that is selected is regarded as being valid and as if it were selected before data inspection (Longford, 2005). The model selection process is far from ignorable.

Maximum likelihood is efficient only asymptotically, and only with a valid model. If simulations confirm that the asymptotic sampling variance closely approximates the sampling variance in a finite sample setting and the bias is small or none, we cannot conclude that maximum likelihood estimation is efficient also in small samples. In finite samples, *some* submodels of a valid model may yield more efficient estimators for *some* targets; the bias of an estimator that is based on an invalid model may be more than offset by the variance reduction in relation to a valid model. This issue is highly relevant in semiparametric models in which the effective numbers of observations and parameters cannot be counted straightforwardly.

Xavier de Luna (*Umeå University*) **and Per Johansson** (*Uppsala University and Institute for Labour Market Policy Evaluation, Uppsala*)

The paper presents in a convincing manner a broad class of models for longitudinal and event time data. Our purpose with this comment is to make potential users of these models aware of an important pitfall in the analysis of studies with waiting time to treatment (i.e. the time units have been eligible for treatment). Indeed, ignoring waiting time is not innocuous even in experiments with randomized treatment assignment. When the outcome of interest is a survival time, then the effect of the treatment can be defined by using the survival functions of the treated and non-treated subjects over the population of those who are eligible for treatment. The survival function $S(t) = E\{\mathcal{I}(T \geqslant t)\}$, where $\mathcal{I}(\cdot)$ is the indicator function, is then estimated for treated and control units respectively. By randomizing treatment assignment to units, waiting time to treatment is balanced for. In general, we would expect the hazard to death to be a function of waiting time $w: h(t; w) = E\{\mathcal{I}(T = t) | T \geqslant t, W = w\}$. Marginalizing over waiting time is equivalent to considering $h(t) = E_W\{h(t; W)\}$. Unfortunately, the use of the estimated average (over w) hazards to construct the Kaplan–Meier estimator does not yield a valid estimator of the population survival function $S(t)$, unless $h(t; w) = h(t)$. This is because the Kaplan–Meier estimator is a non-linear function of the hazards.

In the colon cancer study of Section 5.2, waiting time is taken into account through the covariate Z_{2i}. We must assume that the hazards that are modelled are not functions of w within the classes $Z_{2i} = 0$ ($w \leqslant 20$ days) and $Z_{2i} = 1$ ($w > 20$ days), for the curves that are displayed in Fig. 3 to be interpretable as survival functions. Note that the waiting time is at most 1 month (Fleming, 1992), and this might be a reasonable assumption here.

In non-randomized experiments, controls have no well-defined waiting time. This constitutes a major complication because waiting time cannot then be introduced as a covariate in a model. This issue was addressed in Fredriksson and Johansson (2004) and de Luna and Johansson (2007) by conditioning the inference on waiting time. For instance, the models that are discussed by Zeng and Lin may be applied on a stratum that is defined by a waiting time w_0. Then, in this stratum, all units having survived until time w_0 and not treated at that time can and must be considered as controls. For this approach to be feasible, enough observed cases for each waiting time stratum of interest must be available.

Ross L. Prentice (*Fred Hutchinson Cancer Research Center and University of Washington, Seattle*)
I congratulate the authors on a lucid and impressive paper that unifies and extends a diverse statistical literature, while using a modern empirical process for asymptotic distributional results, including that of semiparametric efficiency.

The authors recommendation (a) is to 'use the new class of transformation models to analyse failure time data'. They motivate this new class of models (2)–(4) by the need to allow for the possibility of crossing hazards. However, as noted in the first sentence of their paper, the Cox (1972) model includes time varying covariates, which may readily be defined to include crossing hazards; see Prentice *et al.* (2005) for a recent, practically important, example. More generally, the Cox model is by far the most important special case of models (2)–(4), because of the ready interpretation of regression effects on the hazard ratio, to the point that I wonder whether the larger class adds much. Linear transformation models as a class do not seem to share such a useful interpretation, except for accelerated failure time models which, as the authors note, are not encompassed by models (2)–(4).

The authors' recommendation (b) is to 'make routine use of random-effects models for multivariate failure time data'. Although frailty models allow for dependence between clustered failure times, and I agree that normal random effects have the advantage of avoiding restrictions on pairwise dependences, I do not find the frailty approach to be so appealing. For example, frailty models typically imply complicated marginal distributions for failure times, and the interpretation of regression coefficients is conditional on the frailty. Why should the marginal models for a failure time change, just because some possibly correlated failure times are being simultaneously analysed? Copula models preserve marginal distributions while also allowing correlation. A multivariate normal copula model as applied to standard normal variates arising from Cox model margins (e.g. Li and Lin (2006)) seems particularly appealing. For recurrent events, the authors argue that the inclusion of post-randomization time-dependent variables in Cox models may affect treatment effect interpretation. However, random-effect modelling cannot be expected to remove biases if censoring rates depend in a complex fashion on the preceding counting process history. Careful data analysis is then required, with Cox models having evolving covariates providing a useful and interpretable modelling context (e.g. Kalbfleisch and Prentice (2002), chapter 9).

I again congratulate the authors on their stimulating work.

N. I. Ramesh (*University of Greenwich, London*) **and A. C. Davison** (*Ecole Polytechnique Fédérale de Lausanne*)
We would like to mention work on a related topic for which the methods that are outlined in this interesting paper may be applicable.

We use a multistate Markov model to analyse the movement of ticks of the species *Ixodes ricinus* up and down blades of grass under the influence of covariates such as temperature, relative humidity or light. The movements are recorded under controlled conditions in a laboratory setting over 10 days, with light changed to mimic diurnal variation (Perret *et al.*, 2003).

A simple model is that at any time a tick is in one of the three states, resting at the foot of the blade (1), walking between the top and bottom of the blade (2) and questing for prey at the top of the blade (3), and that transitions may take place between these as follows: $1 \leftrightarrow 2 \leftrightarrow 3$, so direct transition between states 1 and 3 is not allowed. Under a proportional hazards model, we might suppose that a tick in state 1 moves out of it at a rate $h_{12}(t)\,\xi_{12}(x;\beta)$, where $h_{12}(t)$ depends on time t since the start of the experiment, and represents a base-line rate at which a tick in state 1 might leave that state. Similarly we define base-line rates $h_{32}(t)$, $h_{21}(t)$ and $h_{23}(t)$ for the other possible transitions. The quantities $\xi_{ij}(x;\beta)$ reflect the influence of covariates on the rates.

The key aspect of interest is what influences changes between the states?

This is an application where the diurnal variation would lead us to expect cyclic behaviour, so we may use a proportional hazard model with periodic base-line hazard (Pons and de Turckheim, 1988).

To what extent do the ideas of the present paper extend to the periodic case, or to other situations where a base-line function has some special form or constraints induced by the sampling plan?

Peter Sasieni (*Queen Mary, University of London*)

This impressive paper unifies many models for right-censored data (including right-censored repeated measures) and provides a general approach to asymptotic theory and computation. Although the power of empirical process theory is not doubted, the loss of the key concept of 'history', which is central to martingale theory, is lamentable. More particularly, are these large unifying models simply too big to be useful? If model (4) were widely used, how would one perform a meta-analysis based on published results? The following issues all relate to parameter identifiability, interpretation and approximation (McCullagh, 2002).

(a) In the transformation model (1), the scale is fixed by the error distribution. Since different distributions have different variances, the magnitude of the parameter β has no common interpretation. Would constraining the variance of ε help interpretation of β (Chen *et al.*, 2002a; McCullagh, 2002)?

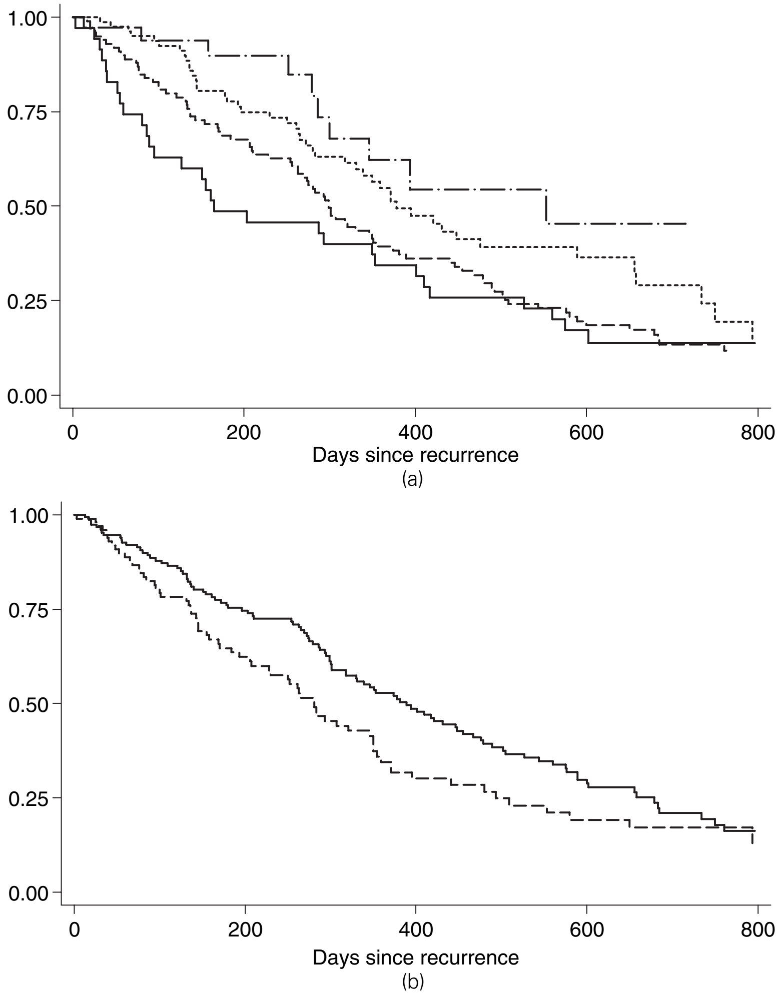

Fig. 8. Kaplan–Meier estimates of survival from cancer recurrence to death in the colon cancer example (a) showing how the survival from recurrence is shorter in those who had recurrence earlier (———, <121 days; – – –, 121–365 days; -------, 366–730 days; · — ·, >730 days) and (b) showing how the survival from recurrence is worse in those in the treatment arm (although overall survival is better owing to the much bigger beneficial effect of treatment on recurrence) (———, treat = 0; – – –, treat = 1)

(b) In general, since the interpretation of the regression parameters β depends on the transformation H, there are difficulties in interpreting confidence intervals that take into account the uncertainty in H (Bickel and Doksum, 1981; Hinkley and Runger, 1984). There are exceptions: notably, when the error is extreme value or logistic, β has a natural interpretation that does not depend on H.
(c) In model (4), one cannot even interpret the sign of β without taking into account γ and Λ. In such circumstances the focus will turn to estimating functionals such as $\Pr(T_1 < T_2|Z_1, Z_2)$ or $E(T_2 - T_1|Z_1, Z_2)$. What then is the advantage of model (4) over non-parametric estimation?
(d) Is model (7) simply too big? As illustrated by Fig. 2, the likelihood can be quite flat with multiple local maxima. In such circumstances we might prefer a suboptimal model with a simpler interpretation that captures the main features of the data. A hierarchy of models would be useful in practice.
(e) Given that the likelihood is not (even asymptotically) convex, can we be sure that the solution to the score equation that is used in the M-step will lead to a consistent estimator?
(f) The model that was used to analyse the colon cancer data treats cancer and death symmetrically despite the fact that there were no cases of death without recurrence. It is of clinical interest to note that the time from recurrence to death is
  (i) strongly correlated with the time to recurrence (Fig. 8) and
  (ii) *decreased* by treatment.

L. Tian (*Northwestern University, Evanston*) **and L. J. Wei** (*Harvard School of Public Health, Boston*)
We thank Professor Zeng and Professor Lin for providing us practically useful, theoretically justifiable inference procedures for a general class of semiparametric models via the maximum likelihood principle. Their proposals are far reaching and can handle various classical challenging problems in survival and longitudinal data analysis. When the fitted model is correctly specified, the resulting estimation procedure is asymptotically efficient. Moreover, unlike other *ad hoc* methods dealing with censored data, theirs is valid without much restriction on the distribution of the censoring variable. The authors also showed that the non-parametric maximum likelihood estimator (NPMLE) is numerically tractable at least when the number of observed failure times is not too large. An alternative way to obtain an approximation to the distribution of the NPMLE of the Euclidean parameter θ_0 is to draw random samples repeatedly from the density function proportional to $\exp\{-\mathrm{pl}_n(\theta)\}$, where $\mathrm{pl}_n(\cdot)$ is the profile log-likelihood function. The realizations of these random samples can be generated via a Markov chain Monte Carlo sampler (Lee *et al.*, 2005).

The authors concluded that 'there is no reason, theoretical or numerical, not to use maximum likelihood estimation for semiparametric regression models'. However, a fitted model is probably an approximation to the true model. It is possible that the NPMLE may not converge under a working model. Moreover, generally the robust 'sandwich variance estimate' for the NPMLE is difficult to obtain owing to a lack of an explicit score function. Furthermore, for evaluating a working model, first we fit the data with the model and then validate it by using, for example, a function D, which measures the average distance between the observed and the model-based predicted responses. Preferably this distance function can be easily interpretable, say, with respect to the scale of the response variable. For example, we may use an R^2-type measure or the absolute prediction error as a possible candidate for D (Uno *et al.*, 2007; Tian *et al.*, 2007). We then use the sampling distribution of an estimated D to evaluate the adequacy of the fitted model. However, a likelihood-based validation criterion may not be easy to interpret. Therefore, to make a coherent package from model estimation to validation, we may prefer to use certain moment-based estimates for the regression parameters.

Lastly, we may take the above approach to compare different working models, e.g. to examine whether a normal random-effects model is better than a gamma frailty counterpart or a parametric model is better than a semiparametric counterpart.

Keming Yu and Shanchao Yang (*Brunel University, Uxbridge, and Guangxi Normal University, Guilin*) **and Ali Gannoun** (*Conservatoire National des Arts et Métiers, Paris*)
This is an impressive piece of work; the idea may motivate some new research for quantile regression in survival analysis. Whereas a conditional survival function at time t represents the proportion of those conditionally surviving up to time t, a pth $(0 < p < 1)$ conditional quantile function provides the earliest time by which the proportion p have died. Let $Q_{h(T)}(p|Z, \tilde{Z})$ be the pth conditional quantile of $h(T)$. First, although it is difficult to find a parametric transformation to achieve normality for a standard linear mixed model or to leave the unknown transformation function under a semiparametric version (10), quantile regression has the feature of equivalance to monotone transformation, i.e. $Q_{h(T)}(p|Z, \tilde{Z}) = h\{Q_T(p|Z, \tilde{Z})\}$ for any

monotone function h, so we can simply select a transformation such as the Box–Cox transformation to apply. In fact, whatever the parametric monotone transformation h to achieve normality of the response variable, we shall transform it back to obtain $Q_T(p|Z, \tilde{Z}) = h^{-1}\{Q_{h(T)}(p|Z, \tilde{Z})\}$. Second, in the context of this paper we may propose semiparametric quantile models in different stages. For example, corresponding to the heteroscedastic version of linear transformation models $H(T) = -\beta^{\mathrm{T}} Z + \exp(-\gamma^{\mathrm{T}} \tilde{Z})\varepsilon$, we may have their quantile versions as $Q_{h(T)}(p|Z, \tilde{Z}) = -\beta(p)^{\mathrm{T}} Z + \exp\{-\gamma(p)^{\mathrm{T}} \tilde{Z}\}$ (Chaudhuri *et al.*, 1997; Koenker and Geling, 2001), where parameters $\beta(p)$ and $\gamma(p)$ depend on p (the monotonicity of $Q_{h(T)}(p|Z, \tilde{Z})$ over p may be required). Corresponding to those new cumulative intensity functions in equations (3) or (4), the new quantile regression models can be derived as follows: from

$$\log[-\log\{P(T > t|Z)\}] = \beta^{\mathrm{T}} Z(s) + \gamma^{\mathrm{T}} \tilde{Z} + \exp(\gamma^{\mathrm{T}} \tilde{Z} - 1)\left[1 + \int_0^t \exp\{\beta^{\mathrm{T}} Z(s)\}\, \mathrm{d}\Lambda(s)\right]\Lambda_0(t),$$

where $\lambda(t)$ is the base-line hazard function and $\Lambda_0(t) = \int_0^t \lambda(s)\, \mathrm{d}s$, we have that $Q_{h(T)}(p|Z, \tilde{Z})$ satisfies the equation

$$Q_{h(T)}(p|Z, \tilde{Z}) = \Lambda_0^{-1}\left(\frac{[\log\{-\log(1-p)\} - \beta^{\mathrm{T}} Z(s) - \gamma^{\mathrm{T}} \tilde{Z}]\exp(1 - \gamma^{\mathrm{T}} \tilde{Z})}{1 + \displaystyle\int_0^{Q_{h(T)}(p|Z, \tilde{Z})} \exp\{\beta^{\mathrm{T}} Z(s)\}\, \mathrm{d}\Lambda(s)}\right).$$

Then $Q_{h(T)}(p|Z, \tilde{Z})$ could be estimated on the basis of the 'check function' or 'loss function' $\rho(u) = u\{p - I(u < 0)\}$ (Portnoy, 2003; Gannoun and Yu, 2007) with proper non-linear optimization.

The **authors** replied later, in writing, as follows.

We are delighted with the unusually large number of contributions from such a diverse group of researchers. We thank all the discussants for taking the time to read our paper and to prepare constructive comments. We are particularly grateful to those who travelled from outside the UK to attend the Ordinary Meeting. For brevity, it is not feasible to respond to all the points that were raised. We shall focus on some common themes.

This year marks the 35th anniversary of Cox's (1972) landmark paper on proportional hazards regression, which is the foundation of our work. We are deeply honoured by Sir David Cox's participation in the meeting and the discussion. As always, his comments are extremely insightful and pertinent. We share his views on hazard and accelerated failure time modelling, contrast between proportionality and additivity, and general *versus* specific models.

Several discussants, particularly Prentice and Sasieni, question the usefulness of transformation models. We wish to reiterate that we are not advocating abandonment of the Cox model, but rather extension of this highly useful model to provide additional modelling capabilities. Although we have presented our models in very general and somewhat abstract forms, any specific application will probably involve only a subset of the models with simpler representation. Most studies are concerned with single events with time invariant covariates, in which setting the class of linear transformation models that is given in equation (1) or its heteroscedastic version that is given above equation (3) would suffice. Although cast within the framework of transformation models, the paper contains new development for the Cox regression, such as Cox models with non-gamma random effects and the joint modelling of repeated measures and failure times via the semiparametric linear mixed model and the Cox model with normal random effects.

There seems to be a general agreement that the proportional hazards assumption should be challenged in practice. Several discussants, including Cox, Breslow, Farewell and Tom, Fine, Kalbfleisch and Xu, and Prentice, recommend adjustment of non-proportionality through the use of manufactured time varying covariates in the form of $Z\, f(t)$, where $f(t)$ is a known function, such as t or $\log(t)$. This approach can be quite useful, especially if we wish to stay within the hazard modelling framework, but it is rather restrictive and data driven. Finding the right form of f can be challenging, particularly when there are multiple continuous covariates. If a linear transformation model, such as the proportional odds model, truly captures the non-proportionality, then that model would provide more concise summarization of the data than the Cox model with manufactured time varying covariates. As Breslow points out, the log-hazard ratio takes the form of $a + b\, \log(t)$ under the two-sample Cox model with $f(t) = \log(t)$ and the form of $a + b\, \log\{\Lambda(t)\}$

under model (3). The latter formulation is actually more appealing since it is non-parametric and scale invariant.

The familiar linear model form of equation (1) is more intuitive than the hazard formulation, especially when the response variable does not pertain to failure time. The choice of the extreme value distribution for ε yields the proportional hazards model. If the true error distribution is not extreme value, then we should use whatever the true distribution is rather than abandoning this attractive formulation.

Equation (1) can be expressed as $g\{S_z(t)\} = H(t) + \beta^{\mathrm{T}} Z$, where $S_z(\cdot)$ is the conditional survival function of T given Z, and $g(\cdot)$ is a known link function. The choices of $g(x) = \log\{-\log(x)\}$ and $g(x) = \log\{x/(1-x)\}$ yield the proportional hazards and proportional odds models respectively. These two models are analogous to the binary data regression models with the complementary log–log- and logit link functions. If the true link function is logit, it would not be sensible to insist on using the complementary log–log-link function and trying to correct for the misspecification of the link function by incorporating interaction terms.

Both equation (1) and its survival function representation show that linear transformation models characterize directly the effects of covariates on the ultimate outcome, i.e. survival time or survival probability. By contrast, hazard is a conditional concept and the effects of covariates on the survival time under the Cox model with (manufactured) time-dependent covariates is not transparent. A few discussants, particularly Henderson, Martinussen and Scheike, Farewell and Tom, and Prentice, are concerned that the effects of covariates on the hazard function may not be clear under transformation models. But we should not confine ourselves to a hazard interpretation, especially when the hazards are not proportional and alternative formulations lead to more parsimonious models.

Hutton remarks that accelerated failure time models are more useful than proportional hazards models in certain medical applications and can accommodate non-proportional hazards through appropriate choices of the error distribution. Her comments further support the use of linear transformation models, which are the same as parametric accelerated failure time models except that the transformation of the failure time is unspecified.

Several discussants, particularly Breslow, Fine and Farrington, Hocine and Moreau, query the interpretation of the regression parameters in the gastric cancer example. In that example, model (3) takes a simple form

$$H(T) = -\beta Z + \exp(-\gamma Z)\varepsilon,$$

where ε has the extreme value distribution. This is just a heteroscedastic linear regression model. The model can also be written in terms of the cumulative hazard function

$$\Lambda(t|Z) = \{\exp(\beta Z)\,\Lambda_0(t)\}^{\exp(\gamma Z)},$$

which is a semiparametric version of the Weibull regression model in which β and γ represent the effects of the combination therapy on the scale and shape of the failure time distribution respectively. Under model (4), the hazard ratio is

$$\exp(\beta+\gamma)\{1+\exp(\beta)\,\Lambda_0(t)\}^{\exp(\gamma)-1},$$

which reduces to equation (14) in Henderson's contribution under $\Lambda_0(t) = t$. As explained by Henderson, the interpretation of β and γ is fairly straightforward in this case. We agree with Hutton and Farrington, Hocine and Moreau that it would be desirable to identify factors that cause crossing hazards.

For recurrent events and time varying covariates, it is necessary to formulate the transformation models in terms of hazard. Then the interpretation of the regression parameters indeed may not be simple, and the main advantages of the transformation models may lie in prediction. Regression analysis has traditionally been focused on individual regression parameters. Inference on individual regression parameters can be quite misleading when covariates are correlated. More emphasis should be placed on prediction, i.e. on characterizing how different covariates act together to affect the ultimate outcomes. For purposes of prediction, it is desirable to use the most accurate model. For that, rich classes of models such as those presented in our paper are highly valuable, as alluded by Cox.

As Sasieni and Li point out, the interpretation of the regression parameters and variance components generally depend on the transformation function. In the special cases of the proportional hazards and proportional odds models, the regression parameters have simple interpretation. Thus, we recommend the use of those two models as long as they provide reasonable approximations.

Henderson and Fine mention the Cox model with time varying regression coefficients. This is a nice way of visualizing the effects of covariates on the hazard function over time. It is, however, very difficult to estimate the time varying coefficients well. Indeed, such parameters cannot be estimated at the usual $n^{1/2}$-rate and non-parametric smoothing is required. The proportional odds model with time varying regression coefficients that was suggested by Martinussen and Scheike encounters the same difficulties. Additive hazards models with time varying regression coefficients are easier to deal with; however, additive models do not constrain the hazard functions to be positive, as noted by Cox.

The advantages of the non-parametric maximum likelihood estimators (NPMLEs) have been discussed at great length in the paper, and we shall not repeat our arguments. As Borgan points out, partial likelihood provides a simple, although inefficient, approach to analysing nested case–control data. This approach applies only to the Cox model and is intractable when covariates are measured with error, whereas the NPMLE is more broadly applicable. As demonstrated by Zeng *et al.* (2006), NPMLEs provide a unified framework for efficient semiparametric inference under the nested case–control and case–cohort designs.

Fine and Prentice find marginal models conceptually more appealing than random-effects models. This view is not universally adopted. Indeed, random-effects models are more desirable than marginal models when the response for an individual rather than for the population is the focus (Zeger *et al.*, 1988). The advantages of random-effects models over marginal models are discussed in Section 7 of our paper.

We share Scheike and Martinussen's sentiment that one should use the random-effect distribution that provides the best description of the correlation. Unfortunately, it is difficult to determine the true distribution of random effects empirically. The advantages of the normal random effects are described in the paper. Normal random effects are particularly natural in joint modelling of repeated measures and failure times.

We agree with Aalen that counting process martingale theory is a very important conceptual framework for formulating the effects of potentially time varying covariates on event history under general censoring mechanisms. Our remarks about the limitations of this tool pertain only to the proofs of asymptotic results. As Aalen points out, counting process martingale theory provides a very simple approach to understanding the properties of standard survival analysis methods, such as Kaplan–Meier and Nelson–Aalen estimators, log-rank tests and Cox regression analysis. Thus, most graduate courses in survival analysis theory are currently taught from the counting process martingale point of view. Once the students start working on their theses, however, they realize that this elegant theory cannot be used to solve cutting edge research problems whereas empirical process theory is much more powerful.

Andersen is right that the construction of the likelihoods makes the standard assumption of independent and non-informative censoring. The theoretical results in the paper cover random left truncation under suitable regularity conditions. In the presence of internal time varying covariates, likelihoods given in expressions (5) and (8) are indeed only partial likelihoods and the EM algorithms may not apply. Like the standard Cox regression analysis, our methods require that time varying covariates be measured at all observed event times.

Since the scope of our paper is very broad, it is impossible to cite all relevant papers. Kosorok *et al.* (2004) is discussed in remark 3 of the paper. The *h*-likelihood that was mentioned by Nelder, Lee and Ha is very intriguing. It is unclear whether this approach will provide numerically accurate and statistically efficient estimators in the semiparametric setting that is considered in our paper.

Tsodikov provides very nice insights into the convergence properties of the semiparametric EM algorithms that are employed in our paper. The robust variance scoring algorithm that was mentioned by Commenges is promising and worth trying. It would also be worthwhile to explore the profile sampler and piggyback bootstrap procedures that were mentioned by Kosorok as well as Tian and Wei.

Bickel and Henderson bring up the issue of estimating the transformation parameter. This issue is briefly discussed in Section 7 and is carefully studied in Zeng and Lin (2007), which shows that the transformation parameter can be estimated reliably from the data and the variability of the estimator can be properly accounted for.

One potential use of the transformation models is to test the proportional hazards model since the latter is embedded in the former. Specifically, we can check the proportional hazards assumption by testing $\rho=0$ under the Box–Cox transformation or by testing $\gamma=0$ under model (3) or (4) with $G(x)=x$. This can be done by the Wald, score or likelihood ratio statistics. The score statistics that were proposed by Philipson and Ho make the unnecessary assumption that β and Λ_0 are known. As Farrington, Hocine and Moreau point out, Quantin *et al.* (1996) proposed a score statistic to test $\gamma=0$ in model (3) with $G(x)=x$. Their statistic does not seem to account properly for the variability due to the estimation of the cumulative base-line hazard function.

Fig. 2 seems to have confused Sasieni as well as Kalbfleisch and Xu. This figure is actually a concatenation of four separate plots for the four classes of bivariate transformation models. For each class of models, the likelihood appears to be convex.

We are pleased to see the interesting extension of our work to doubly censored data that was presented by Chen and Ying and the application of the transformation models to quantile regression that was described by Yu and Yang. It should be possible to apply our theory to the problem that was described by Cuzick. By imposing appropriate constraints on the jump sizes of $\Lambda(\cdot)$ to reflect periodicity, our results should also be applicable to the problem that was outlined by Ramesh and Davison. Fine as well as Farrington, Hocine and Moreau suggest the use of different timescales, such as the gap times between successive events. By formulating the dependence of the gap times through random effects, our framework can cover such models. The non-parametric transformation model that was mentioned by Bickel and the non-parametric random-effects distribution that was mentioned by Horowitz are very challenging problems, and it is unclear whether the NPMLE is feasible in either case.

References in the discussion

Aalen, O. O. (1980) A model for non-parametric regression analysis of counting processes. *Lect. Notes Statist.*, **2**, 1–25.

Aalen, O. O. (1988) Heterogeneity in survival analysis. *Statist. Med.*, **7**, 1121–1137.

Aalen, O. O. (1992) Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Ann. Appl. Probab.*, **2**, 951–972.

Aalen, O. O. (1994) Effects of frailty in survival analysis. *Statist. Meth. Med. Res.*, **3**, 227–243.

Aalen, O. O. and Gjessing, H. K. (2001) Understanding the shape of the hazard rate: a process point of view. *Statist. Sci.*, **16**, 1–22.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical Models based on Counting Processes*. New York: Springer.

Andersen, P. K., Klein, J. P., Knudsen, K. and Palacios, R. T. (1997) Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, **53**, 1475–1484.

Aranda-Ordaz, F. J. (1983) An extension of the proportional hazards model for grouped data. *Biometrics*, **39**, 109–117.

Bagdonavicius, V., Hafdi, M. A. and Nikulin, M. (2004) Analysis of survival data with cross-effects of survival functions. *Biostatistics*, **5**, 415–425.

Bagdonavicius, V. B. and Nikulin, M. S. (1999) Generalized proportional hazards model based on modified partial likelihood. *Liftime Data Anal.*, **5**, 329–350.

Barker, P. and Henderson, R. (2005) Small sample bias in the gamma frailty model for univariate survival. *Liftime Data Anal.*, **11**, 265–284.

Bickel, P. J. (1975) One-step Huber estimates in the linear model. *J. Am. Statist. Ass.*, **70**, 428–434.

Bickel, P. J. and Doksum, K. A. (1981) An analysis of transformations revisited. *J. Am. Statist. Ass.*, **76**, 296–311.

Borgan, Ø., Fiaccone, R. L., Henderson, R. and Barreto, M. L. (2007) Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. *Scand. J. Statist.*, **34**, 53–69.

Borgan, Ø., Goldstein, L. and Langholz, B. (1995) Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.*, **23**, 1749–1778.

Breslow, N. E. and Wellner, J. A. (2007) Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.*, **34**, 86–102.

Bunea, F. and McKeague, I. W. (2005) Covariate selection for semiparametric hazard function regression models. *J. Multiv. Anal.*, **92**, 186–204.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997) On average derivative quantile regression. *Ann. Statist.*, **25**, 715–744.

Chen, G., Lockhart, R. A. and Stephens, M. A. (2002a) Box-Cox transformations in linear models: large sample theory and tests of normality (with discussion). *Can. J. Statist.*, **30**, 177–234.

Chen, K., Jin, Z. and Ying, Z. (2002b) Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**, 659–668.

Commenges, D., Jacqmin-Gadda, H., Proust, C. and Guedj, J. (2006) A Newton-like algorithm for likelihood maximization: the robust-variance scoring algorithm. *Preprint*. (Available from `http://arxiv.org/abs/math/0610402`.)

Commenges, D., Joly, P., Gégout-Petit, A. and Liquet, B. (2007) *Scand. J. Statist.*, **34**, doi:10.1111/j.l467-9469.2006.00536.x.

Cowling, B. J., Hutton, J. L. and Shaw, J. E. H. (2006) Joint modelling of event counts and survival times. *Appl. Statist.*, **55**, 31–39.

Cowling, B. J., Shaw, J. E. H., Hutton, J. L. and Marson, A. G. (2007) New statistical method for analysing time to first seizure: example using data comparing carbamazepine and valproate monotherapy. *Epilepsia*, **48**, 1173–1178.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist.* B, **34**, 187–220.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.

Cuzick, J., Edwards, R. and Segnan, N. (1997) Adjusting for non-compliance and contamination in randomised clinical trials. *Statist. Med.*, **16**, 1017–1029.

Cuzick, J., Sasieni, P., Myles, J. and Tyrer, J. (2007) Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. *J. R. Statist. Soc.* B, **69**, 565–588.

Diggle, P., Liang, K. Y. and Zeger, S. (1994) *Analysis of Longitudinal Data*. Oxford: Clarendon.

Dixon, J. R., Kosorok, M. R. and Lee, B. L. (2005) Functional inference in semi-parametric models using the piggyback bootstrap. *Ann. Inst. Statist. Math.*, **57**, 255–277.

Duchateau, L., Janssen, P., Kezic, I. and Fortpied, C. (2003) Evolution of recurrent asthma event rate over time in frailty models. *Appl. Statist.*, **52**, 355–363.

Elbers, C. and Ridder, G. (1982) True and spurious duration dependence: the identifiability of the proportional hazard model. *Rev. Econ. Stud.*, **49**, 403–409.

Farewell, D. M. (2006) Linear models for censored data. *PhD Thesis*. Lancaster University, Lancaster.

Farrington, C. P. and Whitaker, H. J. (2006) Semiparametric analysis of case series data (with discussion). *Appl. Statist.*, **55**, 553–594.

Feller, W. (1991) *An Introduction to Probability Theory and Its Applications*, vol. 2. New York: Wiley.

Fleming, T. R. (1992) Evaluating therapeutic interventions: some issues and experiences (with discussion). *Statist. Sci.*, **7**, 428–456.

Fosen, J., Borgan, Ø., Weedon-Fekjaer, H. and Aalen, O. O. (2006) Dynamic analysis of recurrent event data using the additive hazard model. *Biometr. J.*, **48**, 381–398.

Fredriksson, P. and Johansson, P. (2004) Dynamic treatment assignment—the consequences for evaluations using observational data. *Discussion Paper 1062*. Institute for the Study of Labor, Bonn.

Gannoun, A., Saracoo, J. and Yu, K. (2007) Comparison of nonparametric estimators of conditional distribution function and quantile regression under censoring for survival analysis. *Statist. Modllng*, to be published.

Grambsch, P. M. and Therneau, T. M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–526.

Ha, I. D. and Lee, Y. (2005) Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models. *Biometrika*, **92**, 717–723.

Ha, I. D. and Lee, Y. (2007) On likelihood estimation in semiparametric frailty models. To be published.

Ha, I. D., Lee, Y. and Song, J.-K. (2001) Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233–243.

Heagerty, P. J. and Zeger, S. L. (2000) Marginalized multilevel models and likelihood inference. *Statist. Sci.*, **15**, 1–26.

Heckman, J. J. and Singer, B. (1984a) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320.

Heckman, J. J. and Singer, B. (1984b) The identifiability of the proportional hazard model. *Rev. Econ. Stud.*, **51**, 231–243.

Hedekar, D. and Gibbons, R. (1994) A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933–944.

Hinkley, D. V. and Runger, G. (1984) The analysis of transformed data (with discussion). *J. Am. Statist. Ass.*, **79**, 302–320.

Honoré, B. (1990) Simple estimation of a duration model with unobserved heterogeneity. *Econometrica*, **58**, 453–473.

Horowitz, J. L. (1998) *Semiparametric Methods in Economics*. New York: Springer.

Horowitz, J. L. (1999) Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica*, **67**, 1001–1028.

Horowitz, J. L. and Lee, S. (2004) Semiparametric estimation of a panel data proportional hazards model with fixed effects. *J. Econometr.*, **119**, 155–198.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. New York: Springer.

Hsieh, F. (2001) On heteroscedastic hazards regression models: theory and application. *J. R. Statist. Soc.* B, **63**, 63–79.

Hutton, J. L. and Monaghan, P. F. (2002) Choice of accelerated life and proportional hazards models for survival data: asymptotic results. *Liftime Data Anal.*, **8**, 375–393.

Hutton, J. L. and Pharoah, P. O. D. (2002) Effects of cognitive, sensory and motor impairment on the survival of people with cerebral palsy. *Arch. Dis. Chldhd*, **86**, 84–89.

Kalbfleisch, J. D. and Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edn. Hoboken: Wiley.

Keiding, N. (1998) Selection effects and nonproportional hazards in survival models and models for repeated events. *In Proc. 19th Int. Biometric Conf.*, pp. 241–250. Cape Town: International Biometric Society.

Koenker, R. and Geling, O. (2001) Reappraising medfly longevity: a quantile regression survival analysis. *J. Am. Statist. Ass.*, **96**, 458–468.

Kosorok, M. R., Lee, B. L. and Fine, J. P. (2004) Robust inference for proportional hazards univariate frailty regression models. *Ann. Statist.*, **32**, 1448–1491.

Langholz, B. (2007) Use of cohort information in the design and analysis of case-control studies. *Scand. J. Statist.*, **34**, 120–136.

Lee, B. L., Kosorok, M. R. and Fine, J. P. (2005) The profile sampler. *J. Am. Statist. Ass.*, **100**, 960–969.

Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006) *Generalised Linear Models with Random Effects: Unified Analysis via h-likelihood*. London: Chapman and Hall.

Li, Y. and Lin, X. (2006) Semiparametric normal transformation models for spatially correlated survival data. *J. Am. Statist. Ass.*, **101**, 591–603.

Longford, N. T. (2005) Model selection and efficiency: is 'Which model . . .?' the right question? *J. R. Statist. Soc.* A, **168**, 469–472.

de Luna, X. and Johansson, P. (2007) Matching estimators for the effect of a treatment on survival times. *Working Paper 2007:1*. Institute for Labour Market Policy Evaluation, Uppsala.

Ma, S. and Kosorok, M. R. (2005) Penalized log-likelihood estimation for partly linear transformation models with current status data. *Ann. Statist.*, **33**, 2256–2290.

Martinussen, T. and Scheike, T. H. (2006) *Dynamic Regression Models for Survival Data*. New York: Springer.

Martinussen, T. and Scheike, T. H. (2007) Aalen additive hazards change-point model. *Biometrika*, to be published.

McCullagh, P. (2002) What is a statistical model? *Ann. Statist.*, **30**, 1225–1310.

McKeague, I. W. and Sasieni, P. D. (1994) A partly parametric additive risk model. *Biometrika*, **81**, 501–514.

Moger, T. A., Aalen, O. O., Heimdal, K. and Gjessing, H. K. (2004) Analysis of testicular cancer data by means of a frailty model with familial dependence. *Statist. Med.*, **23**, 617–632.

Murphy, S. A. (1994) Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.*, **22**, 712–731.

Murphy, S. A. (1995) Asymptotic theory for the frailty model. *Ann. Statist.*, **23**, 182–198.

Murphy, S. A. and van der Vaart, A. W. (2000) On profile likelihood. *J. Am. Statist. Ass.*, **95**, 449–465.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, **19**, 25–44.

Noh, M., Ha, I. D. and Lee, Y. (2006) Dispersion frailty models and HGLMs. *Statist. Med.*, **25**, 1341–1354.

Oakes, D. (1989) Bivariate survival models induced by frailties. *J. Am. Statist. Ass.*, **84**, 487–493.

O'Sullivan, F. (1988) Fast computation of fully automated log-density and log-hazard estimators.. *SIAM J. Sci. Statist. Comput.*, **9**, 363–379.

Perret, J.-L., Guerin, P. M., Diehl, P. A., Vlimant, M. and Gern, L. (2003) Darkness induces mobility, and saturation deficit limits questing duration, in the tick *Ixodes ricinus*. *J. Exptl Biol.*, **206**, 1809–1815.

Pons, O. and de Turckheim, E. (1988) Cox's periodic regression model. *Ann. Statist.*, **16**, 678–693.

Portnoy, S. (2003) Censored regression quantiles. *J. Am. Statist. Ass.*, **98**, 1001–1012.

Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotcher, J., Kuller, L., Limacher, M. and Wactowski-Wende, J. (2005) Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am. J. Epidem.*, **162**, 1–11.

Quantin, C., Moreau, T., Asselain, B., Maccario, J. and Lellouch, J. (1996) A regression survival model for testing the proportional hazards hypothesis. *Biometrics*, **52**, 874–885.

Ridder, G. (1990) The non-parametric identification of generalized accelerated failure time models. *Rev. Econ. Stud.*, **57**, 167–182.

Ridder, G. and Woutersen, T. M. (2003) The singularity of the information matrix of the mixed proportional hazard model. *Econometrica*, **71**, 1579–1589.

Robins, J. M., Hernán, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

Rondeau, V., Commenges, D. and Joly, P. (2003) Maximum penalized likelihood estimation in a gamma frailty model. *Liftime Data Anal.*, **9**, 139–153.

Samuelsen, S. O., Ånestad, H. and Skrondal, A. (2007) Stratified case-cohort analysis of general cohort sampling designs. *Scand. J. Statist.*, **34**, 103–119.

Sasieni, P. D. and Winnett, A. (2003) Martingale difference residuals as a diagnostic tool for the Cox model. *Biometrika*, **90**, 899–912.

Scharfstein, D. O., Tsiatis, A. A. and Gilbert, P. B. (1998) Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Liftime Data Anal.*, **4**, 355–391.

Scheike, T. H. (2006) A flexible semiparametric transformation model for survival data. *Liftime Data Anal.*, **12**, 461–480.

Scheike, T. H. and Juul, A. (2004) Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics*, **5**, 193–206.

Slud, E. V. and Vonta, F. (2004) Consistency of the NPML estimator in the right-censored transformation model. *Scand. J. Statist.*, **31**, 21–41.

Solomon, P. J. (1984) Effect of misspecification of regression models in the analysis of survival data. *Biometrika*, **71**, 291–298.

Struthers, C. A. and Kalbfleisch, J. D. (1986) Mispecified proportional hazards models. *Biometrika*, **73**, 363–369.

Tetens, J. N. (1786) *Einleitung zur Berechnung der Leibrenten udt Anwartschaften II*. Leipzig: Weidmanns Erben und Reich.

Therneau, T. M. and Grambsch, P. M. (2000) *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L. J. (2007) Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, to be published.

Tsiatis, A. A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statist. Sin.*, **14**, 793–818.

Tsodikov, A. (2003) Semiparametric models: a generalized self-consistency approach. *J. R. Statist. Soc.* B, **65**, 759–774.

Uno, H., Cai, T., Tian, L. and Wei, L. J. (2007) Evaluating prediction rules for *t*-year survivors with censored regression model. *J. Am. Statist. Ass.*, to be published.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1998) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zeng, D. and Lin, D. Y. (2007) Semiparametric transformation models with random effects for recurrent events. *J. Am. Statist. Ass.*, **102**, 167–180.

Zeng, D., Lin, D. Y., Avery, C. L., North, K. E. and Bray, M. S. (2006) Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics*, **7**, 486–502.