STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2535

A shared random effects model for censored medical costs and mortality

Lei Liu*,†, Robert A. Wolfe and John D. Kalbfleisch

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.

SUMMARY

In this paper, we propose a model for medical costs recorded at regular time intervals, e.g. every month, as repeated measures in the presence of a terminating event, such as death. Prior models have related monthly medical costs to time since entry, with extra costs at the final observations at the time of death. Our joint model for monthly medical costs and survival time incorporates two important new features. First, medical cost and survival may be correlated because more 'frail' patients tend to accumulate medical costs faster and die earlier. A joint random effects model is proposed to account for the correlation between medical costs and survival by a shared random effect. Second, monthly medical costs usually increase during the time period prior to death because of the intensive care for dying patients. We present a method for estimating the pattern of cost prior to death, which is applicable if the pattern can be characterized as an additive effect that is limited to a fixed time interval, say b units of time before death. This 'turn back time' method for censored observations censors cost data b units of time before the actual censoring time, while keeping the actual censoring time for the survival data. Time-dependent covariates can be included. Maximum likelihood estimation and inference are carried out through a Monte Carlo EM algorithm with a Metropolis–Hastings sampler in the E-step. An analysis of monthly outpatient EPO medical cost data for dialysis patients is presented to illustrate the proposed methods. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: proportional hazards model; mixed model; re-censoring; change point; informative censoring; survival analysis

1. INTRODUCTION

Medical cost data are collected routinely by hospitals, disease registries, and health insurance companies. The statistical analysis of medical cost data has gained increasing interest recently. For example, models for medical cost data provide estimates that can be used in

*Correspondence to: Lei Liu, Department of Public Health Sciences, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, VA 22908-0717, U.S.A.
†E-mail: liulei@virginia.edu

Contract/grant sponsor: Centers for Medicare & Medicaid Services (CMS); contract/grant number: 500-96-0007

Received 14 October 2004
Accepted 17 January 2006

cost-effectiveness analyses, which in turn inform policy makers to maximize health benefits for individuals and society.

The statistical analysis of medical cost data at the patient level requires flexible models in order to capture the complex relationship between costs and other events (such as death) during the follow-up period. However, the analysis of medical cost and survival time is often complicated by censoring due to incomplete follow-up where neither the survival time nor the remaining medical cost record is observed after censoring. For example, standard right-censored analysis methods (e.g., the Kaplan–Meier estimator) of the cumulative lifetime medical cost are invalid because the lifetime cost is not independently censored even when there is independent censoring of the survival time [1].

A variety of methods have been developed to analyse medical cost data solely. Lin *et al.* [1] studied the non-parametric estimator for the mean total cost. They divided the study time period into a small fixed number of intervals, then estimated the mean total cost by summing up the sample mean of the total costs from those observed to die within each interval multiplied by the Kaplan–Meier estimate for death probability in that interval. Cost history data can be used to derive the mean total cost in a similar way. Zhao and Tsiatis [2] formulated the survival function of quality-adjusted lifetime (QAL) which, akin to medical cost, is a random transformation of survival time. They used the inverse probability of censoring weighting technique (IPCW [3]) to obtain an unbiased non-parametric estimate of survival function of QAL. Similarly Bang and Tsiatis [4] developed a non-parametric estimate for lifetime cost by IPCW technique and studied its efficiency. Their partitioned estimator was later simplified by Jiang and Zhou [5] with a bootstrap confidence interval proposed for the mean of medical costs.

Lin [6] proposed a proportional means model for complete cost history data, but it required that the censoring time is known or is completely random. Lin [7] also developed a linear regression model which assumes additive covariate effects on the mean medical cost. IPCW method was used to correct the bias induced by informative censoring of the medical cost. Both papers targeted the semi-parametric modelling of marginal mean/rate of total medical cost, while the relationship between survival time and cost is not specified. They made rather restrictive assumptions on the relationship between lifetime cost and covariate information. Jain and Strawderman [8] enhanced the Hazard Regression (HARE) model [9] with IPCW technique, resulting in a more flexible model for the lifetime cost distribution. Lin [10] summarized the regression analysis of incomplete medical cost data. He also proposed a pattern mixture approach which models the conditional means of cost accumulation given specific survival patterns.

The joint modelling of survival and medical cost is often important, as for example, in studies of cost-effectiveness. Fine and Gelber [11] proposed a joint regression analysis of survival and QAL and estimated the parameters by U-statistics. Huang [12] devised a calibration regression model for survival time and lifetime medical cost. Both papers assume a linear covariate effect on survival time and QAL/lifetime medical cost in a semi-parametric fashion with unspecified bivariate distribution for the error terms.

In this paper, we proposed a joint model for monthly medical costs and survival time that simultaneously accounts for several features that have not been accounted for in previous models. Specifically, the medical costs history data are treated as repeated measures in the presence of death and censoring. The model accounts for patterns of cost related to both time measured since entry and to time measured relative to death. The model can include time-dependent covariates, allowing estimates of the costs associated with transient effects. The model also

accounts for potential correlation between monthly medical costs and survival time, which could occur if more 'frail' patients tend to have a higher death rate as well as a larger monthly medical cost.

The problem of joint modelling of repeated measures and time-to-event data has been addressed by several authors, e.g., References [13–18]. In our paper a shared random effect is incorporated in the model to induce the correlation among costs in different months for each patient in addition to that between costs and survival time.

Another feature in our model allows an additive component in medical costs to be present during a known fixed time interval before death. Such patterns have often been observed in medical practice but are typically not fully studied in published statistical papers. For example, Lin *et al.* [1] and Bang and Tsiatis [4] both considered a death cost but the extra cost was limited to the final year or month (which could overlap two intervals) prior to death.

Faucett *et al.* [19] and Pauler and Finkelstein [20] also considered a change point in the joint modelling of repeated measures and time-to-event data, both in the Bayesian setting. They assumed a parametric distribution for the change point measured as time from entry. However, the patterns of medical costs often appear to be more closely linked to the death time than to the entry time. Descriptively, our model accounts for the pattern of the cost change in time measured retrospectively from death rather than prospectively from entry.

In the next section, we present the joint random effects model. In Section 3, we develop corresponding estimation methods with an EM algorithm. A method is proposed that links costs to both time since entry and time prior to death. We then assess the operating characteristics of the proposed inference procedures by simulations in Section 4. In Section 5, we apply our method to medical cost data of kidney patients on dialysis. Concluding remarks are given in Section 6.

2. MODEL

Let C_i and D_i be the independent censoring and death times for subject i $(i=1,2,\ldots,n)$, respectively. Write $X_i=\min(C_i,D_i)$ as the follow-up time and $\Delta_i=I(D_i\leqslant C_i)$, where $I(\cdot)$ is the indicator function. Let $Y_i(t)=I(X_i\geqslant t)$ be the at-risk indicator. Denote by $N_i^D(t)=I(X_i\leqslant t,\Delta_i=1), 0<t<\infty$, the counting process for the death process. The death hazard at time t is $\lambda(t)$. Denote by $U_i(j)$ the incremental medical cost (possibly transformed, e.g. log transformation) accumulated during the jth month, where $j=1,2,\ldots,n_i$, with $n_i=\lfloor X_i\rfloor$ being the integer month death time for subject i, where $\lfloor\ \rfloor$ is the floor function. We ignore the medical cost for the final partial month by taking the floor of X_i for convenience. Define by constant b the known time ahead of death when the medical cost changes. Let v_i be the random effect with a parametric distribution which affects both the medical cost and the survival rate. Define by $Z_i(t)$ the covariate vector (possible external time-dependent) at time t. We assume both $Z_i(\cdot)$ and $U_i(\cdot)$ change only at discrete time point $1,2,\ldots,n_i$, thus $Z_i(t)=Z_i(j)$ for $t\in[j-1,j)$. Assume censoring is independent of death time. The joint model of medical cost and death for subject i is written as

$$U_i(j)=\mu_j+\beta^{\mathrm{T}}Z_i(j)+v_i+f(j,D_i,\eta,b)+e_{ij} \tag{1}$$

$$\lambda_i(t)=\lambda_0(t)\exp(\alpha^{\mathrm{T}}Z_i(t)+\gamma v_i) \tag{2}$$

where β, α, γ and η are unknown parameters, respectively. μ_j is the unknown baseline monthly medical cost, which can be used to model medical cost pattern after entry. For simplicity $f(j, D_i, \eta, b)$ is assumed to be a known function and depends on j and D_i only through $D_i - j$, which is the time axis relative to death. As an example, for step rise with jump size η, $f(j, D_i, \eta, b) = \eta I(0 < D_i - j < b)$ for the medical cost after change point and $f(j, D_i, \eta, b) = 0$ if $D_i - j > b$. It may also involve random effects, as shown in the application. Write $\underline{U_i} = \{U_i(1), U_i(2), \ldots, U_i(n_i)\}$ as the observed medical cost history vector up to month n_i. Assume $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ are independent of (v_i, C_i, D_i). The 'likelihood' for the data $\mathbf{O}_i \equiv \{\underline{U_i}, X_i, \Delta_i\}$ given random effect v_i is

$$\begin{aligned} L(\underline{U_i}, X_i, \Delta_i | v_i) &= \prod_{j=1}^{n_i} L(U_i(j) | X_i, \Delta_i, v_i) L(X_i, \Delta_i | v_i) \\ &= \frac{1}{(\sqrt{2\pi}\sigma_e)^{n_i}} \exp\left[-\frac{1}{2\sigma_e^2} \sum_{j=1}^{n_i} e_{ij}^2 \right] [\lambda_0(x_i) \exp(\alpha^{\mathrm{T}} Z_i(x_i) + \gamma v_i)]^{\Delta_i} \\ &\quad \times \exp\left[-\int_0^{x_i} \exp(\alpha^{\mathrm{T}} Z_i(t) + \gamma v_i) \, \mathrm{d}\Lambda_0(t) \right] \end{aligned} \tag{3}$$

where x_i is the realization of X_i and

$$e_{ij} = U_i(j) - \mu_j - v_i - \beta^{\mathrm{T}} Z_i(j) - f(j, D_i, \eta, b)$$

Note that some of e_{ij}'s are not known because D_i is unobserved for the censored subjects.

3. ESTIMATION AND INFERENCE

The estimation of parameters in (3) is complicated because some of the factors cannot be evaluated due to the unknown D_i for censored subjects. Integrating D_i out of (3) is algebraically complicated and does not have a closed form. Furthermore, the empirical survival function cannot be integrated when the last death event is censored and the death hazard is undefined after that time without further parametric assumption on the baseline hazard $\lambda_0(t)$. To handle this dilemma, we propose a simple 'turn back time' or 're-censoring' method for the medical cost of censored observations, which excludes the unknown factors in the likelihood.

We can move back the censoring time for medical cost by b months so we only use the cost history $\{U_i(j), C_i - j > b\}$ for censored subjects. From (3), we exclude factors for $\{U_i(j), C_i - j \leqslant b\}$ for censored subjects, which involve the unobserved value of $f(j, D_i, \eta, b)$. Note we keep the censoring time for the survival model (2) unchanged. Write $n_i^* = \Delta_i n_i + (1 - \Delta_i)(\lfloor X_i \rfloor - b)$ as the new number of follow-up months for medical cost. Similarly, we write the new censored medical cost vector as $\underline{U_i^*}$. Then we can obtain the joint log-likelihood for the new observed data $\mathbf{O}_i^* \equiv \{\underline{U_i^*}, X_i, \Delta_i\}$ and v_i as

$$\begin{aligned} l^* &\equiv \log L(\underline{U_i^*}, X_i, \Delta_i, v_i) \\ &= \sum_{j=1}^{n_i^*} [\log L(U_i(j) | X_i, \Delta_i, v_i) + \log L(X_i, \Delta_i | v_i) + \log p(v_i)] \end{aligned} \tag{4}$$

where $p(v_i)$ is the density function of v_i. The detailed form can be obtained by replacing n_i by n_i^* in the corresponding terms of (3).

Next we will show that the estimating equation for the likelihood (4) is unbiased despite the exclusion of part of the medical cost data. We only consider the parameters in cost model since the survival data remain the same for the parameter estimates in (3) with the usual definition of independent censoring. As an example, the score equation for β in (4) given v can be written as

$$\frac{\partial l^*}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial l_i^*}{\partial \beta} = \frac{1}{\sigma_e^2} \sum_{i=1}^{n} \sum_{j=1}^{n_i^*} e_{ij} Z_i(j) \tag{5}$$

To demonstrate that the conditional expectation of (5) is 0, we have

$$\begin{aligned} E\left(\frac{\partial l_i^*}{\partial \beta} \middle| v_i \right) &= \frac{1}{\sigma_e^2} E_{X,\Delta} E\left[\sum_{j=1}^{n_i^*} e_{ij} Z_i(j) | X_i, \Delta_i, v_i \right] \\ &= \frac{1}{\sigma_e^2} E_{X,\Delta} \sum_{j=1}^{n_i^*} E[e_{ij} Z_i(j) | X_i, \Delta_i, v_i] \\ &= 0 \end{aligned}$$

The result follows since v_i, $\Delta_i = I(D_i \leqslant C_i)$ and $\lfloor X_i \rfloor$, thereby n_i^*, are all independent of e_{ij}. The unbiasedness of other parameters in the cost model can be justified similarly.

We assume that $v_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_v^2)$. Since v_i's are unobserved, it is natural to use EM algorithm to obtain the MLE for parameters $\theta \equiv \{\alpha, \beta, \gamma, \eta, \mu_j, \sigma_e^2, \sigma_v^2\}$. In the M-step we can easily take the first and second derivatives of l^* with respect to θ. For completeness, we report all score components and second partial derivatives in Appendix A.

Since there is no closed form for the density of $f(v_i | \mathbf{O}_i^*)$ in the E-step, Metropolis–Hastings (M–H) algorithm (Appendix B) can be used to generate M random numbers $v_i^{(m)}$ $(m=1,\ldots,M)$ for the estimation of the expectation of the sufficient statistics involving frailties. Examples are $\hat{E}(v_i | \mathbf{O}_i^*) = (1/M) \sum_{m=1}^{M} v_i^{(m)}$ and $\hat{E}(\exp(\gamma v_i) | \mathbf{O}_i^*) = (1/M) \sum_{m=1}^{M} \exp(\gamma v_i^{(m)})$.

We use Louis' formula [21] to obtain the information matrix for observed data likelihood. The observed information matrix $I(\hat{\theta})$ is

$$I(\hat{\theta}) = -\hat{E}\left\{ \frac{\partial^2 l^*}{\partial \theta \, \partial \theta'} \middle| \mathbf{O}^*, \hat{\theta} \right\} - \hat{E}\left\{ \frac{\partial l^*}{\partial \theta} \frac{\partial l^*}{\partial \theta'} \middle| \mathbf{O}^*, \hat{\theta} \right\} + \hat{E}\left\{ \frac{\partial l^*}{\partial \theta} \middle| \mathbf{O}^*, \hat{\theta} \right\} \hat{E}\left\{ \frac{\partial l^*}{\partial \theta'} \middle| \mathbf{O}^*, \hat{\theta} \right\}$$

All three terms are evaluated at the last iteration of the EM algorithm, when the last term becomes zero for the MLE $\hat{\theta}$. The first two expectations can be calculated by averaging over the corresponding terms involving M–H values.

Ad hoc measures of 'information loss' due to the turn back time method are defined for both β and η. It is defined for β by the percent of months turned back (PML_β, percentage months loss for β), i.e.

$$\mathrm{PML}_\beta = \frac{\sum_{i=1}^{n} (1 - \Delta_i) \min(b, \lfloor C_i \rfloor)}{\sum_{i=1}^{n} \lfloor X_i \rfloor} \times 100 \text{ per cent}$$

The month lost for η results from the number of months turned back after the change point for censored subjects among the total months observed after the change point, i.e.

$$\mathrm{PML}_\eta = \frac{\sum_{i=1}^{n}(1-\Delta_i)\{\lfloor C_i \rfloor - \max(\lfloor D_i \rfloor - b, 0)\}}{\sum_{i=1}^{n}(1-\Delta_i)\{\lfloor C_i \rfloor - \max(\lfloor D_i \rfloor - b, 0)\} + \Delta_i \min(b, \lfloor D_i \rfloor)} \times 100 \text{ per cent}$$

4. SIMULATION

In this section, we conduct simulations under two settings to evaluate the performance of the proposed estimation procedures. In both settings, we consider a single binary covariate Z which takes value 0 or 1 each with probability $\frac{1}{2}$. The sample size $n = 200$. The regression coefficients are $\alpha = 1$ and $\beta = (\beta_0, \beta_1)^{\mathrm{T}} = (0, 1)^{\mathrm{T}}$. $\mu_j = \phi \times j$ with $\phi = 0.2$. γ takes value 0 and 1 in setting I and II, respectively. Both v_i's and e_{ij}'s are generated from independent normal distribution with mean 0 and variance 1 ($\sigma_v^2 = \sigma_e^2 = 1$). The baseline intensity function for death is taken to be exponential with constant 0.05. We assume there is a step rise of monthly medical costs 3 months ($b = 3$) before death, i.e. $f(j, D_i, \eta, b) = \eta I(0 < D_i - j < b)$. The value of η is taken to be 1. Since death time is not in integer month, we ignore the medical cost for the last partial month, e.g. if a subject dies at 8.5 month, the medical cost will jump at the beginning of the fifth month. We assume a random censoring time $C = 6 + \mathrm{U}(0, 6)$ for all subjects. Six hundred replicates are generated for each setting.

In setting I ($\gamma = 0$), each subject has on average 5.7 monthly medical cost records. The censoring rate is 47 per cent. Rolling back medical cost data for censored subjects affects about 25 per cent months (PML_β) of all observed months, which represent 12 per cent (PML_η) of the month loss for η. For setting II ($\gamma = 1$), subjects have 5.4 monthly medical cost records on average and 46 per cent of subjects are censored. About 11 per cent of the monthly medical cost records with respect to η are lost due to the re-censoring and 26 per cent are lost for estimation of β.

We summarize the simulation results in Tables I and II, respectively. It can be seen that the magnitudes of the empirical biases of estimates from the shared random effects model are

Table I. Simulation results: parameter estimates for setting I.

	Shared random effects model				Separate marginal models			
Parameter	Bias	SE	SEM	CP (%)	Bias	SE	SEM	CP (%)
$\beta_0 = 0$	0.003	0.132	0.131	94.2	0.003	0.132	0.130	94.3
$\beta_1 = 1.0$	−0.009	0.184	0.181	93.3	−0.010	0.176	0.174	94.5
$\phi = 0.2$	−0.001	0.021	0.022	96.8	−0.001	0.020	0.021	94.6
$\eta = 1.0$	0.004	0.144	0.145	95.3	0.003	0.122	0.119	93.8
$\alpha = 1.0$	0.019	0.222	0.210	93.2	0.010	0.219	0.209	93.0
$\gamma = 0$	−0.001	0.152	0.155	96.3				
$\sigma_e^2 = 1.0$	0.004	0.056	0.055	95.0	0.004	0.056	0.055	95.0
$\sigma_v^2 = 1.0$	0.002	0.135	0.135	95.3	−0.007	0.135	0.133	94.0

Bias is the mean of the parameter estimates (based on 600 replicates) minus the true value; SE is the standard error of the parameter estimates; SEM is the sampling mean of the standard error estimate; CP is the coverage probability of the corresponding 95 per cent confidence interval.

Table II. Simulation results: parameter estimates for setting II.

	Shared random effects model				Separate marginal models			
Parameter	Bias	SE	SEM	CP (%)	Bias	SE	SEM	CP (%)
$\beta_0 = 0$	−0.001	0.131	0.138	95.8	−0.102	0.125	0.120	85.8
$\beta_1 = 1.0$	−0.018	0.172	0.184	95.2	−0.238	0.151	0.154	64.7
$\phi = 0.2$	−0.002	0.021	0.023	96.2	−0.041	0.020	0.021	48.2
$\eta = 1.0$	0.023	0.140	0.153	96.7	0.478	0.127	0.123	4.0
$\alpha = 1.0$	0.015	0.291	0.286	96.0	−0.239	0.204	0.201	75.3
$\gamma = 1.0$	0.010	0.233	0.235	96.3				
$\sigma_e^2 = 1.0$	0.005	0.057	0.057	96.5	0.023	0.058	0.058	95.2
$\sigma_v^2 = 1.0$	−0.016	0.173	0.185	94.2	−0.297	0.106	0.109	27.8

very small in both settings. The coverage probabilities are close to the nominal level 0.95. We also observe only minor biases for variance estimates.

For comparison, we use separate marginal models for estimation assuming that medical cost and death hazard share no random effect, i.e. $\gamma = 0$. We fit the medical cost data by the conventional linear mixed model and survival data by a standard proportional hazards model separately. The results are shown in the right side of each table. The separate marginal models are correctly specified in setting I but incorrectly in setting II when the dependence between medical costs and terminal event through the shared random effect is present. As seen from the tables, the separate marginal model parameter estimates in setting I are unbiased, but they are severely biased in setting II. The coverage probabilities are very poor, especially for η in setting II. It is clear that ignoring the dependence can result in significant biases, as in the joint modelling of repeated measures and informative drop-out [16, 17].

We also compare the results of the shared random effects model and the separate marginal models in setting I. The biases in both model are very close and the shared random effects model has little increase in variance. In conclusion, these two models have the same accuracy and precision in the special case of $\gamma = 0$.

In both settings, the distributions for the parameter estimates are approximately symmetric and normal (histograms not shown). Figure 1 gives the estimates of cumulative baseline death hazard functions for months $1, 2, \ldots, 10$. We draw the true cumulative baseline hazard functions $\Lambda_0(t) = 0.05t$ for comparison. It can be seen that $\hat{\Lambda}_0(t)$ in both settings is virtually unbiased. We also plot the estimates of $\Lambda_0(t)$ in the marginal model. As expected, $\hat{\Lambda}_0(t)$ in setting I obtained by the separate model is unbiased. The bias in $\hat{\Lambda}_0(t)$ in setting II (Figure 1(d)) arises from the fact that the separate survival model estimates the marginal baseline death hazard rather than the hazard conditional on random effect.

5. APPLICATION

We apply the proposed method to medical cost data extracted from the Medicare outpatient dialysis claims. Erythropoietin (EPO) was prescribed by doctors during most dialysis sessions to improve regulation of patients' red blood cell production. Medicare paid about $1.4 billion in 2002 for outpatient EPO usage [22]. The average outpatient EPO cost per session in each

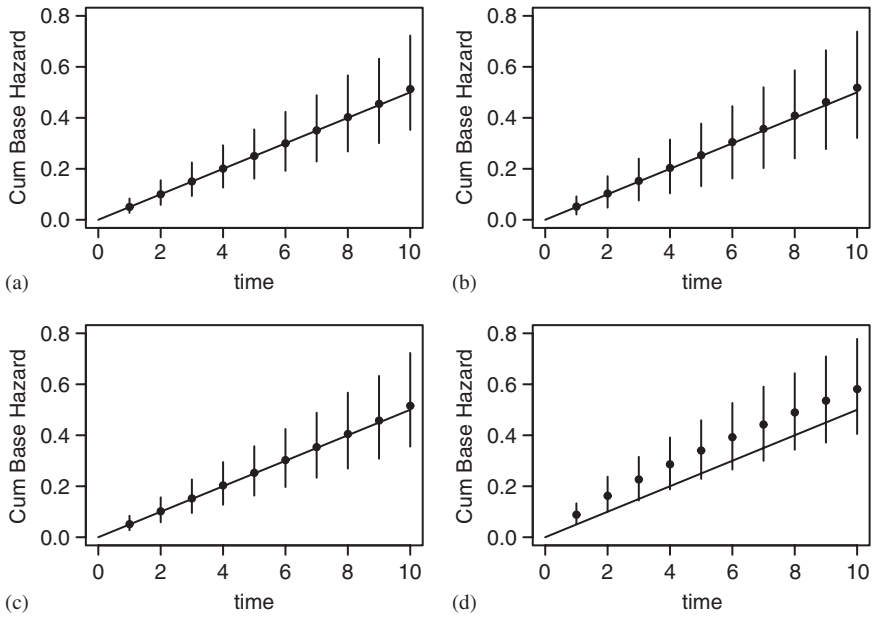

Figure 1. The estimates and 95 per cent confidence intervals for baseline cumulative hazard $\Lambda_0(t)$: (a) shared random effects model at $\gamma = 0$; (b) shared random effects model at $\gamma = 1$; (c) separate marginal model at $\gamma = 0$; and (d) separate marginal model at $\gamma = 1$. Means of the hazard functions at each time point are denoted by dot; pointwise 95 per cent empirical confidence interval for the estimated cumulative baseline hazards are obtained from 600 replicates.

month paid by Medicare is of interest in this study. We suspect that patients with poor health received more EPO prescription per dialysis session, thus incurring more costs. We are interested in joint modelling of monthly outpatient EPO costs and survival (measured continuously in monthly units), taking account of covariate information.

A preliminary study, using a linear regression model, analysed the outpatient EPO payment per session (Erik Roys, personal communication). It showed an increasing pattern in monthly outpatient EPO costs starting from 6 months prior to death. It also exhibited a monthly outpatient EPO cost jump initially since entry time, followed by a linear drop.

In order to have a moderate sample size and to reduce the computational burden, we arbitrarily chose the first 300 patients whose initial dialysis started in July 2000. The follow-up for outpatient EPO cost and survival ended on 31 December 2002. Among them there are 159 males (53 per cent), 216 white (72 per cent), and 133 (44 per cent) patients with diabetes as the primary cause of kidney diseases. The average and median ages at registration are 74 and 75, respectively, and 193 (64 per cent) died during follow-up. Others were censored either before or at the end of study. The mean weight at baseline is 72 kg. The average follow-up is 17.9 months. As before, we ignore the outpatient EPO cost of the last partial month for convenience.

Figure 2 shows the final 8 month trajectories of monthly outpatient EPO costs for 15 randomly selected dead subjects (with follow-up time greater than 16 months to avoid the high-cost initiation period). Many show an increasing pattern which starts around

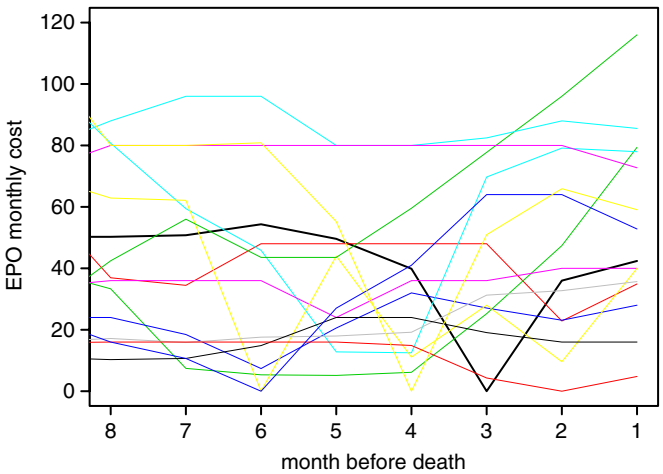

Figure 2. Monthly EPO costs for 15 randomly selected subjects with terminal event.

4–6 months before death. We thus set the change point time as 6 months before death ($b=6$) and re-censored the outpatient EPO cost data for the originally censored patients 6 months prior to their original censoring time. This reduces the average follow-up time to 15.8 months. A covariate 'End' measured time prior to death for this linear pattern (up to and including 6 months). We also created a variable 'Start' to capture the outpatient EPO cost pattern since entry: an initial cost increase in the second month, then a linear decreasing pattern through the eighth month after entry. Months 2–7 were coded numerically, while months 1 and $8,\ldots,30$ were coded with value 8. Other early costs such as high cost accrued after diagnosis as in the treatment of heart disease or cancer, can be similarly incorporated in our model.

From preliminary analysis, we only included Age (in years) and Gender ($1=$male, $0=$female) as predictors for the death hazard. Race, Gender, and Diabetes are not significant for outpatient EPO cost in this sample.

We also noticed from Figure 2 the variation in the slopes of the outpatient EPO cost increase before death. A random slope ω was included in the model to account for this variation. We assume $\omega_i \overset{\text{iid}}{\sim} \text{N}(0,\sigma_\omega^2)$ is independent of both v_i and e_{ij} for $i=1,2,\ldots,n$. The final model is

$$\begin{aligned} U_i(j) &= \beta_0 + \beta_1\,\text{Age} + \beta_2\,\text{Weight} + \beta_3\,\text{Start} + v_i + (\eta+\omega_i)\,\text{End} + e_{ij} \\ \lambda_i(t) &= \exp(\alpha_1\,\text{Age} + \alpha_2\,\text{Gender} + \gamma v_i)\lambda_0(t) \end{aligned} \tag{6}$$

As summarized in Table III, we find that Age and Weight are significant for outpatient EPO cost. There is a significant linear increasing pattern in monthly outpatient EPO cost before death ($\hat{\eta}=1.65$, $p<0.0001$). A linear decreasing pattern for the outpatient EPO cost starting from the second month to eighth month after entry is also highly significant ($\hat{\beta}_3=-1.50$, $p<0.0001$). Age is significant for survival. Each 1 year increase in age elevates the death hazard by 2.7 per cent. The estimate of γ, which models the correlation between monthly outpatient EPO cost and survival, is 0.0099, which is highly significant ($p=0.0008$), suggesting that death hazard is higher for patients with larger random effect

Table III. Analysis of EPO cost data for kidney patients.

	Shared random effects model			Separate marginal models		
	Estimate	SE	p-value	Estimate	SE	p-value
For medical cost						
Intercept	82.5	21.3	0.0001	67.6	24.0	0.005
Age	−0.79	0.30	0.009	−0.45	0.25	0.07
Weight	0.60	0.19	0.001	0.41	0.15	0.006
Start	−1.50	0.16	<0.0001	−1.55	0.16	<0.0001
End	1.65	0.35	<0.0001	1.77	0.37	<0.0001
σ_e^2	587	16		541	12	
σ_v^2	787	103		1150	128	
σ_ω^2	6.57	1.83		12.0	2.23	
For survival						
Age	0.027	0.010	0.006	0.032	0.010	0.0008
Gender	0.19	0.15	0.21	0.18	0.15	0.22
γ	0.0099	0.0030	0.0008			

(or higher monthly medical costs). Ignoring this correlation and fitting the data by separate marginal models produces biased estimates, as shown in the right-hand side of Table III. For example, the effects of Age and Weight in the marginal model weaken substantially. In particular, Age is only marginally significant in the marginal model (coefficient $=-0.45$, $p=0.07$), while in the joint model it becomes significant (coefficient $=-0.79$, $p=0.009$).

A random slope is present in both the shared random effects and the separate marginal models as the 95 per cent confidence interval for $\hat{\sigma}_\omega^2$ is 6.6 ± 3.6 in the shared random effects model and 12.0 ± 4.4 in the separate marginal models. In the shared random effects model, the magnitude of $\hat{\sigma}_\omega = 2.56$ relative to $\hat{\eta} = 1.65$ suggests that although the trend in pre-death costs is increasing on average, it varies substantially from person to person. We also notice that random variation for both v and ω is reduced in the shared random effects model.

One aspect of checking the adequacy of the adopted model is to evaluate the estimated cumulative death hazard functions for various stratifications. As an example, we divided the subjects into two age groups: Younger ($\leqslant 75$ years) and Older (>75 years). We fitted the age-stratified models with the same adjusting variables as in (6). A plot of $\log \hat{\Lambda}_0(t)$ *versus* $\log t$ is displayed in Figure 3. The parallelism of the curves suggests that the proportional hazards model for age is a very good approximation, after adjustment for other covariates.

We also show the residual plot for age in Figure 4. There is no apparent pattern in the residual with respect to age. Further model checking techniques such as likelihood ratio tests to compare our model with more complex models can also be used to check the validity of our model.

Dialysis clinicians suggested that outpatient EPO costs might be elevated in months during which patients are hospitalized. To model this, we carried out an additional analysis that included a dichotomous time-dependent indicator of hospitalization in the cost model. The results were virtually unchanged from those reported in Table III and also showed a higher outpatient EPO cost ($4.62, $p<0.0001$) in months of hospitalization, as was suggested by the clinicians.

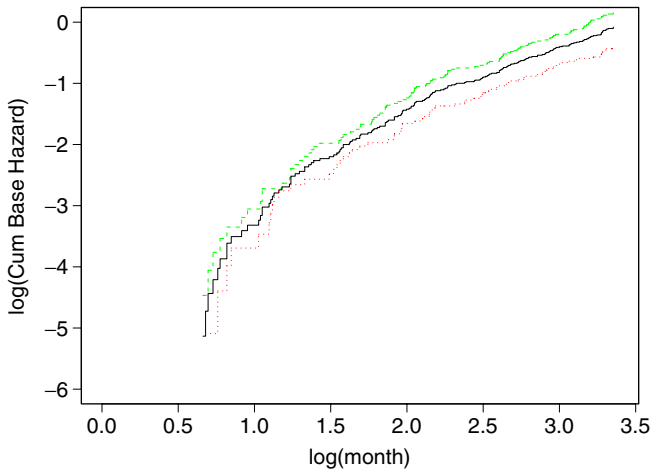

Figure 3. Model checking: cumulative baseline hazard $\log \Lambda_0(t)$ for age effect (—, All; ···, Younger; - - -, Older).

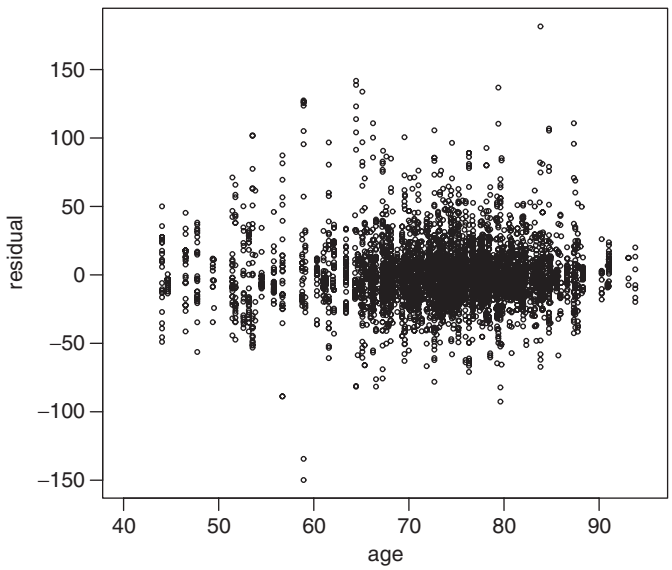

Figure 4. Residual plot for age in EPO cost model.

6. DISCUSSION

In this paper, we propose a shared random effects model for monthly medical costs and survival time. We introduce a simple 'turn back time' method to obtain unbiased parameter estimates for medical cost when there is a cost component during a fixed time interval prior to death. Our simulation shows that the separate marginal models for medical costs and survival can yield very biased results while our models apparently yield consistent results.

Our model is a descriptive model which simultaneously models death and serial cost measures on two time scales: time since entry and time before death. As shown in model (1), it allows us to incorporate cost change pattern both after diagnosis (by μ_j) and before death (by the function $f(\cdot)$), the latter a novel feature to model the effect of time before death on medical cost. The model can also include time-dependent covariates (by $Z_i(j)$), yielding estimates of the costs associated with transient conditions such as hospitalization. The resulting estimates for the joint distribution of costs and mortality will be useful for a variety of descriptive and inference purposes. In the example shown, it brings new insight on the relationship of medical costs and death. It can also be generalized to the joint analysis of death and other positive scaled measures, such as dose or utilization.

Our model is not designed to give results for a single cost effectiveness measure. However, the resulting estimates may be used for more detailed evaluations of the effectiveness of medical care. In particular, the costs associated with death (or entry into care) may be useful for evaluating the effectiveness of medical care given at the end of life (or at the initiation of care). Finally, the estimate resulting from the time-dependent covariate may be useful for evaluating the effectiveness of episodic treatment methods.

Although our original model specifies a change in medical costs at some fixed known time b, in many practical situations only an upper bound on the duration of the exceptional pattern of cost before death is required for 're-censoring'. A variety of models can be used for this pattern, including random effects model in our example, so long as the re-censoring exceeds the maximal duration of the pattern. This upper bound on the change point time can be approximated by preliminary analysis on the uncensored patients. Subject matter knowledge may also help to find the bound. If the bound is chosen too small or the functional form is incorrect, the resulting estimates will be biased. However, if the bound chosen is too large or the functional form for $f(\cdot)$ is overly flexible, there will likely be efficiency losses due to loss of information to re-censoring or due to the large number of parameters in $f(\cdot)$.

In our model, we make a simple assumption that the monthly medical costs can be transformed to have a normal distribution, most often by log transformation. Other transformations, e.g. Box–Cox transformation, may be used for a more thorough investigation. In our application, a portion of the observations have zero cost and non-zero cost observations are highly skewed to the right. Zhou [23] studied the inference on population means for lifetime health care costs with such a non-normal distribution. For convenience, we assume the frailty (e^{v_i}) distribution as log normal. Gamma frailty distribution has been adopted by many other authors. However, Pickles and Crouchley [24] and O'Quigley and Stare [25] showed that for estimation and testing of regression coefficients, it is not critical on the choice of parametric frailty distribution, suggesting the robustness of frailty models. Further model assessment and sensitivity analysis tools can be found in References [26] and [27, Chapter 31] for joint longitudinal and drop-out data. Dobson and Henderson provided various informal graphical diagnostic tools for preliminary model evaluation. Molenberghs and Verbeke also described the local influence approach to identifying the influential subjects. It will be an interesting topic to adapt these tools to our model setting.

Faucett *et al.* [19] used Multiple Imputation technique [28] to reduce the variation in parameter estimates and provide robustness to model mis-specification. Multiple imputation exploits the entire observed medical cost history but may be computationally intensive and involve more complicated models. The 'turn back time' method is simpler but the information loss on medical cost history may be high for heavily censored data. However, in the medical cost

setting, data can be extracted easily with little expenses from large database of observational studies (e.g. hospital records or medicare claims). The information loss may not be a big burden for the analysis in practice.

It is straightforward to generalize our model (1) to the common mixed model form

$$\mathbf{U} = \beta^{\mathrm{T}}\mathbf{Z}_{\mathrm{F}} + v^{\mathrm{T}}\mathbf{Z}_{\mathrm{R}} + \mathbf{f}_{\eta} + \mathbf{e}$$

where $\mathbf{Z}_{\mathrm{F}}$ and $\mathbf{Z}_{\mathrm{R}}$ are covariate vectors (possible time-dependent) for fixed and random effects, respectively. Similarly we can extend (2) to include interaction terms between random effects and other covariates. The corresponding likelihood and estimation equations are readily adapted with only a minor modification for the (expected) functions of random effects. In this paper, we assume a simple covariance structure for the error term e but generalization to more complicated form such as AR(1) is possible, although more computationally demanding. $\mathbf{f}_{\eta}$ may be estimated non-parametrically, e.g. by splines. More complex joint models of repeated measures and event time data may be adapted to our setting as well.

In the estimation process, we use ML instead of REML. These two methods are asymptotically equivalent and ML is relatively easy to implement in our setting. If the sample size is small, REML is preferred to reduce the bias in the variance estimate [29].

We implicitly assume the repeated measures (medical costs) and survival are positively correlated ($\gamma > 0$). But our model works well when they have a negative association (simulation results not shown), which might arise if the repeated measures (e.g. exercise time each week) are protective against the terminal event.

In one analysis, we included hospitalization as a time-dependent covariate in the outpatient EPO cost model, which demonstrates that time-dependent covariate could be analysed. Time-dependent covariates, such as hospitalization, can be interpreted as outcome of treatment. The model with such intermediate outcome as predictors is often useful for explaining mechanisms that lead to variation in costs and mortality. Liu *et al.* [30] proposed a shared frailty model for recurrent events (such as hospitalizations) and a terminal event. A joint model of monthly medical costs, hospitalizations and death time may be of great interest for further study.

The random effect for the slope in the application suggests that the increase in monthly medical costs, prior to the death time, varies among subjects. However, the computation was less stable when the random slope effects were included. Further study on computational algorithms may help to stabilize the estimation. Also, presence of both random effects of v and w makes the computation highly demanding in memory due to the large number of draws by M–H algorithm in the E-step. This is especially a problem in R which is the language we used in carrying out the developmental and illustrative work on our model. It is expected that an implementation with a computing language that used memory more effectively would make it possible to implement the model in a much larger sample.

APPENDIX A

In the M-step, the score equations for $\{\alpha, \gamma, \lambda_0(\cdot), \sigma_e^2, \sigma_v^2\}$ are

$$\frac{\partial l^*}{\partial \alpha} = \sum_{i=1}^{n} \left[Z_i(x_i)\Delta_i - \int_0^{\infty} Y_i(t) Z_i(t) \exp(\alpha^{\mathrm{T}} Z_i(t)) \hat{E}(\exp(v_i\gamma) | \mathbf{O}_i^*) \, \mathrm{d}\Lambda_0(t) \right]$$

$$\frac{\partial l^*}{\partial \gamma} = \sum_{i=1}^{n} \left[\Delta_i \hat{E}(v_i | \mathbf{O}_i^*) - \int_0^\infty Y_i(t) \hat{E}(v_i \exp(v_i\gamma) | \mathbf{O}_i^*) \exp(\alpha^{\mathrm{T}} Z_i(t)) \, \mathrm{d}\Lambda_0(t) \right]$$

$$\frac{\partial l^*}{\partial \lambda_0(x_i)} = \frac{\Delta_i}{\lambda_0(x_i)} - \sum_{k=1}^{n} Y_k(x_i) \exp(\alpha^{\mathrm{T}} Z_k(x_i)) \hat{E}(\exp(v_k\gamma) | \mathbf{O}_k^*)$$

$$\frac{\partial l^*}{\partial \sigma_e^2} = -\frac{N}{2\sigma_e^2} + \sum_{i=1}^{n} \sum_{j=1}^{n_i^*} \frac{e_{ij}^{*2}}{2\sigma_e^4}$$

$$\frac{\partial l^*}{\partial \sigma_v^2} = -\frac{n}{2\sigma_v^2} + \frac{\sum_{i=1}^{n} \hat{E}(v_i^2 | \mathbf{O}_i^*)}{2\sigma_v^4}$$

where $N = \sum_{i=1}^{n} \sum_{j=1}^{n_i^*} 1$ and

$$e_{ij}^* = U_i(j) - \mu_j - E(v_i | \mathbf{O}_i^*) - \beta^{\mathrm{T}} Z_i(j) - f(j, D_i, \eta, b)$$

The Breslow-type baseline hazard estimate for $\lambda_0(\cdot)$ can be written as

$$\hat{\lambda}_0(x_i) = \frac{\Delta_i}{\sum_k Y_k(x_i) \hat{E}(\exp(v_i\gamma) | \mathbf{O}_k^*) \exp(\alpha^{\mathrm{T}} Z_k(x_i))}$$

Denote by $\dot{f}(j, D_i, \eta, b)$ the first derivative with respect to η. For coefficients $\xi = \{\beta, \eta, \mu_j\}$ in model (1), define $\mathbf{x}_{ij} = \{Z_i(j), \dot{f}(j, D_i, \eta, b), \mathbf{i}_j\}^{\mathrm{T}}$ as the overall ijth row of the covariate matrix $\mathbf{X}$, where $\mathbf{i}_j$ is the indicator vector for month with the jth element to be 1 and other elements to be 0. Then we can write the score for ξ as

$$\frac{\partial l^*}{\partial \xi} = \frac{1}{\sigma_e^2} \sum_{i=1}^{n} \sum_{j=1}^{n_i^*} e_{ij}^* \mathbf{x}_{ij}$$

The second derivatives for (α, γ) are

$$\frac{\partial^2 l^*}{\partial \alpha^2} = -\sum_{i=1}^{n} \int_0^\infty Y_i(t) Z_i(t)^{\otimes 2} \exp(\alpha^{\mathrm{T}} Z_i(t)) \hat{E}(\exp(v_i\gamma) | \mathbf{O}_i^*) \, \mathrm{d}\Lambda_0(t)$$

$$\frac{\partial^2 l^*}{\partial \gamma^2} = -\sum_{i=1}^{n} \int_0^\infty Y_i(t) \hat{E}(v_i^2 \exp(v_i\gamma) | \mathbf{O}_i^*) \exp(\alpha^{\mathrm{T}} Z_i(t)) \, \mathrm{d}\Lambda_0(t)$$

and

$$\frac{\partial^2 l^*}{\partial \alpha \partial \gamma} = -\sum_{i=1}^{n} \int_0^\infty Y_i(t) Z_i(t) \hat{E}(v_i \exp(v_i\gamma) | \mathbf{O}_i^*) \exp(\alpha^{\mathrm{T}} Z_i(t)) \, \mathrm{d}\Lambda_0(t)$$

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

More components of the information matrix are given as follows:

$$\frac{\partial^2 l^*}{\partial \xi^2} = -\frac{1}{\sigma_e^2}\mathbf{X}\mathbf{X}^{\mathrm{T}}$$

$$\frac{\partial^2 l^*}{\partial (\sigma_e^2)^2} = \frac{N}{2\sigma_e^4} - \sum_{i=1}^{n}\sum_{j=1}^{n_i^*}\frac{e_{ij}^{*2}}{\sigma_e^6}$$

$$\frac{\partial^2 l^*}{\partial (\sigma_v^2)^2} = \frac{n}{2\sigma_v^4} - \frac{\sum_{i=1}^{n}\hat{E}(v_i^2|\mathbf{O}_i^*)}{\sigma_v^6}$$

$$\frac{\partial^2 l^*}{\partial \lambda_0(x_i)^2} = -\frac{\Delta_i}{\lambda_0(x_i)^2}$$

$$\frac{\partial^2 l^*}{\partial \alpha\, \partial \lambda_0(x_i)} = -S^{(1)}(\alpha, x_i)$$

$$\frac{\partial^2 l^*}{\partial \gamma\, \partial \lambda_0(x_i)} = -S^{(1)}(\gamma, x_i)$$

with

$$S^{(1)}(\alpha, t) = \sum_{k=1}^{n} Y_k(t)\hat{E}(\exp(v_k\gamma)|\mathbf{O}_k^*)Z_k(t)\exp(\alpha^{\mathrm{T}}Z_k(t))$$

and

$$S^{(1)}(\gamma, t) = \sum_{k=1}^{n} Y_k(t)\hat{E}(v_k\exp(v_k\gamma)|\mathbf{O}_k^*)\exp(\alpha^{\mathrm{T}}Z_k(t))$$

All other off-diagonal terms are zero.

APPENDIX B

M–H algorithm is taken to generate the random number chain $v_i^{(m)}$ $(m = 1, \ldots, M)$ of $f(\mathbf{O}_i^*|v_i)$ due to the difficulty of sampling directly from

$$f(v_i|\mathbf{O}_i^*) = \frac{f(\mathbf{O}_i^*|v_i)f(v_i)}{f(\mathbf{O}_i^*)} = \frac{f(\mathbf{O}_i^*|v_i)f(v_i)}{\int f(\mathbf{O}_i^*|v_i)f(v_i)\,\mathrm{d}v_i}$$

At the kth E-step, the M–H chain starts with an initial value $v_i^{(1)}$. Then we can proceed iteratively. After obtaining $v_i^{(m)}$, a new value $\tilde{v}$ is sampled from gamma frailty with variance

$\theta_{(k)}$. An independent random number u is drawn from U(0, 1). $v_i^{(m+1)}$ is obtained as

$$v_i^{(m+1)} = \begin{cases} \tilde{v} & \text{if } u \leqslant \min\left(1, \dfrac{f(\mathbf{O}_i^*|\tilde{v})}{f(\mathbf{O}_i^*|v_i^{(m)})}\right) \\ v_i^{(m)} & \text{otherwise} \end{cases}$$

In the above formula $f(\mathbf{O}_i^*)$ is cancelled in the ratio.

ACKNOWLEDGEMENTS

The authors thank the referees and associate editor for their careful reading. They thank Erik Roys at Kidney Epidemiology and Cost Center of University of Michigan for preparing the medical cost data set. This research was partly supported by the Centers for Medicare & Medicaid Services (CMS) under contract number 500-96-0007. We are also grateful to Drs Debashis Ghosh, Yulei He, and Zhangsheng Yu for helpful discussions.

REFERENCES

1. Lin DY, Etzioni R, Feuer EJ, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997; **53**:419–434.
2. Zhao H, Tsiatis A. A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* 1997; **84**:339–348.
3. Robins J, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology-Methodological Issues*, Jewell N, Dietz K, Farewell V (eds). Birkhauser: Boston, 1992; 297–331.
4. Bang H, Tsiatis A. Estimating medical costs with censored data. *Biometrika* 2000; **87**:329–343.
5. Jiang H, Zhou X. Bootstrap confidence intervals for medical costs with censored observations. *Statistics in Medicine* 2004; **23**:3365–3376.
6. Lin DY. Proportional means regression for censored medical costs. *Biometrics* 2000; **56**:775–778.
7. Lin DY. Linear regression analysis of censored medical costs. *Biostatistics* 2000; **1**:35–47.
8. Jain AK, Strawderman RL. Flexible hazard regression modeling for medical cost data. *Biostatistics* 2002; **3**:101–118.
9. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**:78–94.
10. Lin DY. Regression analysis of incomplete medical cost data. *Statistics in Medicine* 2003; **22**:1181–1200.
11. Fine J, Gelber RD. Joint regression analysis of survival and quality-adjusted survival. *Biometrics* 2001; **57**:376–382.
12. Huang Y. Calibration regression of censored lifetime medical cost. *Journal of the American Statistical Association* 2002; **97**:318–327.
13. De Gruttola V, Tu XM. Modeling Progression of CD4+ lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**:1003–1014.
14. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**:465–480.
15. Hogan J, Laird N. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; **16**:239–257.
16. Wu M, Bailey K. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 1989; **45**:939–955.
17. Wu M, Bailey K. Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine* 1989; **7**:337–346.
18. Xu J, Zeger S. Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* 2001; **50**:375–387.
19. Faucett CL, Schenker N, Taylor JMG. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* 2002; **58**:37–47.
20. Pauler DK, Finkelstein DM. Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes. *Statistics in Medicine* 2002; **21**:3897–3911.
21. Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1982; **44**:226–233.

22. USRDS Annual Data Report, Table K.1. Available on-line at http://www.usrds.org/2004/ref/K_tables_04.pdf
23. Zhou X-H. Inference about population means of health care costs. *Statistical Methods in Medical Research* 2002; **11**:327–339.
24. Pickles A, Crouchley R. A comparison of frailty models for multivariate survival data. *Statistics in Medicine* 1995; **14**:1447–1461.
25. O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. *Statistics in Medicine* 2002; **21**:3219–3233.
26. Dobson A, Henderson R. Diagnostics for joint longitudinal and dropout time modeling. *Biometrics* 2003; **59**:741–751.
27. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. Springer: New York, 2005.
28. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
29. Searle SR, Casella G, McCulloch CE. *Variance Components*. Wiley: New York, 1992.
30. Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics* 2004; **60**:747–756.

