STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:1181–1200 (DOI: 10.1002/sim.1377)

Regression analysis of incomplete medical cost data

D. Y. Lin[*,†]

*Department of Biostatistics, University of North Carolina, CB#7420 McGavran-Greenberg,
Chapel Hill, NC 27599-7420, U.S.A.*

SUMMARY

The accumulation of medical cost over time for each subject is an increasing stochastic process defined up to the instant of death. The stochastic structure of this process is complex. In most applications, the process can only be observed at a limited number of time points. Furthermore, the process is subject to right censoring so that it is unobservable after the censoring time. These special features of the medical cost data, especially the presence of death and censoring, pose major challenges in the construction of plausible statistical models and the development of the corresponding inference procedures. In this paper, we propose several classes of regression models which formulate the effects of possibly time-dependent covariates on the marginal mean of cost accumulation in the presence of death or on the conditional means of cost accumulation given specific survival patterns. We then develop estimating equations for these models by combining the approach of generalized estimating equations for longitudinal data with the inverse probability of censoring weighting technique. The resultant estimators are shown to be consistent and asymptotically normal with simple variance estimators. Simulation studies indicate that the proposed inference procedures behave well in practical situations. An application to data taken from a large cancer study reveals that the Medicare enrollees who are diagnosed with less aggressive ovarian cancer tend to accumulate medical cost at lower rates than those with more aggressive disease, but tend to have higher lifetime costs because they live longer. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: censoring; economic evaluation; generalized estimating equations; health care; inverse probability of censoring weighting; pattern-mixture models.

1. INTRODUCTION

There is tremendous recent interest in the economic evaluation of health care. For instance, many clinical trials now collect data on the costs of treatments. Similar data are routinely collected by hospitals, insurance companies and disease registries. These data provide valuable opportunities to ascertain the cost of treating patients with a particular disease, to compare the costs of alternative intervention/prevention programmes, and to identify determinants of medical cost.

[*] Correspondence to: D. Y. Lin, Department of Biostatistics, University of North Carolina, CB#7420 McGavran-Greenberg, Chapel Hill, NC 27599-7420, U.S.A.

[†] E-mail: lin@bios.unc.edu

Received December 2001
Accepted July 2002

As an example, we consider the linked SEER (Surveillance, Epidemiology and End Results) Medicare database [1], which contains extensive information on 1264345 Medicare enrollees over 65 years old who were diagnosed with cancer from 1973 to 1989. The data on survival time and monthly medical expenditures were collected during the period of 1984 –1990. Detailed clinical, demographic and geographic information was also recorded. A major objective was to determine how the costs of care for these subjects were affected by the type of cancer diagnosed, the clinical stage of the disease, as well as the demographic and geographic characteristics.

There are several complications with the SEER-Medicare database and other available data sources. First, subjects may not survive beyond the time period of interest, and survival time is related to cost accumulation. Secondly, both survival time and cost accumulation process are subject to right censoring. In the SEER-Medicare database, censoring was caused by the limited study duration; in other studies, loss to follow-up is also a major source of censoring. Thirdly, the cost data are normally recorded in broad time intervals, say monthly or yearly intervals, and no information is available on how the cost is accumulated within an interval. In the SEER-Medicare database, the cost data were recorded in monthly intervals. Finally, the costs in different time intervals tend to be correlated. These complications pose major challenges in the statistical analysis of cost data, especially the regression analysis. In fact, it is not even obvious how to construct plausible regression models for such data. The type of censoring encountered here cannot be handled by standard survival analysis methods, as pointed out by Lin *et al.* [2] and elaborated in the next section.

Because of the aforementioned complications, there has been little progress in the development of regression methods for incomplete medical cost data. Lin [3] showed how to perform linear regression on the marginal mean of the total cost. The linear model, however, is not flexible enough for this type of data, as will be discussed in the following. Furthermore, since a subject who dies sooner has less time to accumulate cost, the marginal mean may not fully capture the effects of covariates on cost accumulation, especially if the covariates (such as treatment assignments) have substantial effects on the survival time. Currently, there does not exist any method for separating the effects of covariates on the survival time from the effects of covariates on the rate of cost accumulation.

In the next section, we present generalized linear models for the marginal mean of the total cost and for the conditional means of cost accumulation given specific survival patterns; these models provide great flexibilities in formulating the effects of covariates on various aspects of cost accumulation. We then develop the corresponding estimating equations, which yield consistent and asymptotically normal estimators. In Section 3 we assess the operating characteristics of the proposed inference procedures in practical settings. In Section 4 we provide an application to the aforementioned SEER-Medicare database. Some concluding remarks are given in Section 5.

2. REGRESSION METHODS

2.1. Data structures

Suppose that data are collected on n study subjects over the time period $(0,\tau]$. For $i=1,\ldots,n$, let $Y_i(t)$ be the cumulative cost up to time t. Although $Y_i(.)$ $(i=1,\ldots,n)$ are defined in continuous time, they can only be measured periodically. The potential observation time points

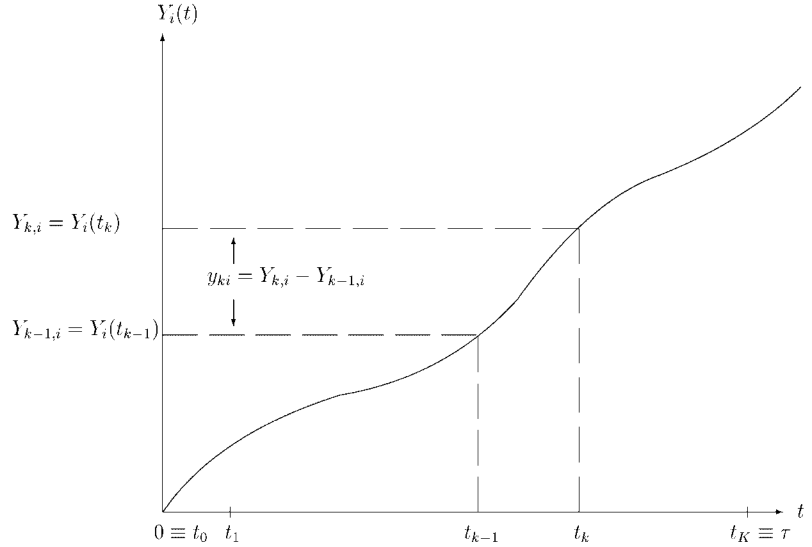

Figure 1. Definitions of time intervals and associated costs.

for $Y_i(.)$ $(i=1,\dots,n)$ are denoted by $t_1 < t_2 < \cdots < t_K \equiv \tau$. Let $Y_{ki} = Y_i(t_k)$ and $y_{ki} = Y_{ki} - Y_{k-1,i}$ with $t_0 = 0$ and $Y_{0,i} = 0$. These definitions are illustrated in Figure 1. It is assumed that K is small relative to n.

Let T_i and C_i be the latent survival time and censoring time for the ith subject, and let $\mathbf{Z}_i(.)$ be the corresponding (possibly time-dependent) covariates. Write $X_i = T_i \wedge C_i$, and $\delta_i = I(T_i \leqslant C_i)$, where $a \wedge b = \min(a,b)$, and $I(.)$ is the indicator function. Naturally, there is no further accumulation of cost after death so that $Y_i(t) = Y_i(t \wedge T_i)$ and $Y_{ki} = Y_i(t_k \wedge T_i)$.

As in the case of $Y_i(.)$, the sample path of $\mathbf{Z}_i(.)$, as a continuous-time process, can seldom be completely observed. Most applications only involve time-independent covariates. To enhance modelling capabilities, we allow covariates to depend on time intervals. Specifically, let $\mathbf{Z}_{ki}$ denote a $p \times 1$ vector of covariates for the ith subject in the kth time interval.

In practice, the population is heterogeneous in that subjects accumulate costs at different rates over time. As illustrated in Figure 2, a subject who accumulates costs at higher rates tend to produce greater cumulative costs at both the censoring time and survival time than a subject with lower accumulation rates, so that there is a positive correlation between $Y_i(C_i)$ and $Y_i(T_i)$ even if C_i is completely independent of T_i and $Y_i(.)$. Because of this induced dependence, standard survival analysis methods, which require independence of the response and its censoring variable, cannot be applied to the censored cost data $\{\tilde{Y}_i, \delta_i, \mathbf{Z}_i(.)\}$ $(i=1,\dots,n)$, where $\tilde{Y}_i = Y_i(T_i) \wedge Y_i(C_i)$ or $Y_i(T_i \wedge C_i)$.

2.2. Regression models

It is crucial to construct models which reflect the underlying physical process and which can be estimated from the available data. One useful class of models is the following generalized linear models for the marginal distributions of the y_{ki}'s:

$$E(y_{ki}|\mathbf{Z}_{ki}) = g(\boldsymbol{\beta}'\mathbf{Z}_{ki}), \quad k=1,\dots,K; \; i=1,\dots,n \tag{1}$$

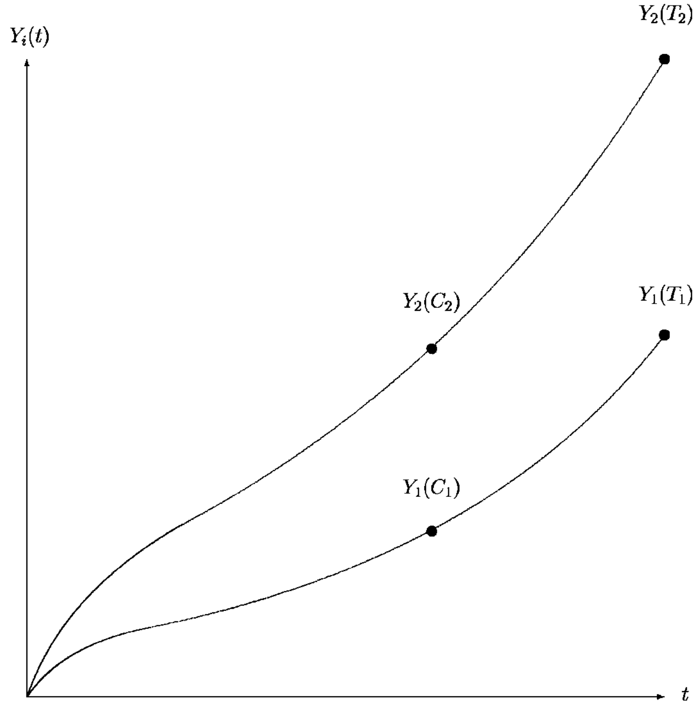

Figure 2. Cumulative costs at survival times and censoring times for two subjects in a heterogeneous population: subject 2 accumulates costs at higher rates than subject 1.

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters, and g is a specific function linking the linear predictor $\boldsymbol{\beta}'\mathbf{Z}_{ki}$ to the marginal mean of y_{ki}. Identity and exponential link functions, $g(x)=x$ and $g(x)=\mathrm{e}^x$, are of particular interest in our context. The choice of g depends on the substantive knowledge as well as the empirical evidence. The marginal means modelled by (1) are highly relevant to public health because they are directly related to the total medical cost in the population. Note that the distributional forms and dependence structures for the y_{ki} $(k=1,\ldots,K)$, that is, the stochastic features of $Y_i(.)$, are completely unspecified. Thus, (1) is semi-parametric.

The formulation in (1) allows the possibilities of identical or different regression effects among the K intervals. To amplify this point, suppose that one is interested in the effects of time-independent covariates $\mathbf{Z}_i$ on cost accumulation. If one defines $\mathbf{Z}_{1i}=(\mathbf{Z}_i',\mathbf{0}',\ldots,\mathbf{0}')',\ldots,\mathbf{Z}_{Ki}=(\mathbf{0}',\ldots,\mathbf{0}',\mathbf{Z}_i',)'$ and $\boldsymbol{\beta}=(\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_K')'$, then $\boldsymbol{\beta}'\mathbf{Z}_{ki}=\boldsymbol{\beta}_k'\mathbf{Z}_i$ $(k=1,\ldots,K)$ so that $\boldsymbol{\beta}_k$ pertains to the effect of $\mathbf{Z}_i$ on medical cost in the kth interval. On the other hand, if one sets $\mathbf{Z}_{1i}=\cdots=\mathbf{Z}_{Ki}=\mathbf{Z}_i$, then $\boldsymbol{\beta}$ is the common effect of $\mathbf{Z}_i$ over the K intervals. Intermediate models can also be created in which a subset of $\mathbf{Z}_i$ has a common effect while the rest does not.

Lin [3] studied the following special case of (1):

$$E(y_{ki}|\mathbf{Z}_i)=\boldsymbol{\beta}_k'\mathbf{Z}_i, \quad k=1,\ldots,K; \; i=1,\ldots,n \tag{2}$$

This model implies that

$$E(Y_{ki}|\mathbf{Z}_i) = \left(\sum_{l=1}^{k} \boldsymbol{\beta}_l\right)' \mathbf{Z}_i, \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{3}$$

Thus, the effects of covariates on the total cost up to τ can be determined. Model (1) is much more general than (2), not only because it allows non-identity link functions, but also because it provides greater flexibilities in parameterizing the covariate effects.

Another important member of (1) is

$$E(y_{ki}|\mathbf{Z}_i) = \mu_k \mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_i}, \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{4}$$

which implies that

$$E(Y_{ki}|\mathbf{Z}_i) = \left(\sum_{l=1}^{k} \mu_l\right) \mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_i}, \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{5}$$

Note that $\mathbf{Z}_i$ includes the constant 1 in (2) and (3), but not in (4) and (5). Models (4) and (5) are reminiscent of the proportional hazards model [4], and are referred to as proportional rates/means models in that the covariates have proportionate effects on the rate/mean of cost accumulation over time. These models and the more general version of multiplicative or log-linear models $E(y_{ki}|\mathbf{Z}_{ki}) = \mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_{ki}}$ $(k=1,\ldots,K;\ i=1,\ldots,n)$ are particularly appealing since the y_{ki}'s are positive.

The fact that the y_{ki}'s are positive suggests that it may be more appropriate to express the y_{ki}'s in (2) on a logarithmic scale. In theory, one may express the y_{ki}'s in (2) or other members of (1) on any scale. Unfortunately, $\sum_{l=1}^{k} y_{li} = Y_{ki}$ only when the y_{ki}'s are defined on the original scale. Furthermore, health care researchers are interested in the mean of the cost on the original scale rather than on a transformed (for example, logarithmic) scale because only the former can be translated into the total cost in the population. Thus, linear models are not as useful as log-linear models for cost data.

The marginal means $E(y_{ki})$ and $E(Y_{ki})$ $(k=1,\ldots,K;\ i=1,\ldots,n)$ pertain to the actual costs incurred by the subjects, incorporating the fact that subjects who die cannot accumulate further cost. Thus, these measures and the associated regression models are highly important, especially from the point of view of public health. The marginal means, however, tend to be lower for the subjects who die sooner, so that an intervention that kills subjects tends to reduce cost. Thus, it is also desirable to assess the effects of covariates on cost accumulation among subjects with similar survival experiences.

To incorporate the survival information into the definition of cost accumulation, we consider $E(y_{ki}|T_i > t_k)$, the mean of the incremental cost in the kth interval $(t_{k-1}, t_k]$ given that the subject survives beyond that interval, and $E(Y_i|t_{k-1} < T_i \leqslant t_k)$, the mean of the lifetime cost $Y_i \equiv Y_i(T_i)$ given that the subject dies in the kth interval $(t_{k-1}, t_k]$. The corresponding generalized linear models take the form

$$E(y_{ki}|T_i > t_k; \mathbf{Z}_{ki}) = g(\boldsymbol{\beta}'\mathbf{Z}_{ki}), \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{6}$$

and

$$E(Y_i|t_{k-1} < T_i \leqslant t_k; \mathbf{Z}_{ki}) = g(\boldsymbol{\beta}'\mathbf{Z}_{ki}), \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{7}$$

Again, the stochastic structures of $Y_i(.)$ are entirely arbitrary, so that the models are semiparametric. Both (6) and (7) are referred to as pattern-mixture models [5]. In general, $\boldsymbol{\beta}$ has different meanings among models (1), (6) and (7); we use the same notation for the sake of simplicity. The above discussion on the choices of g and $\mathbf{Z}_{ki}$ for model (1) also applies to models (6) and (7). The use of models (6) and (7) may provide valuable insights into the underlying mechanisms of cost accumulation because the effects of differential survival times are removed, although such models are not particularly useful in predicting future costs at the baseline since the survival time is not known *a priori*.

2.3. Inference procedures

It is straightforward to make inference under model (6). Clearly, y_{ki} is known if $X_i \geqslant t_k$. Assume that C_i is independent of T_i and $Y_i(.)$ conditional on $\mathbf{Z}_i(.)$, so that

$$E(y_{ki}|X_i \geqslant t_k; \mathbf{Z}_{ki}) = E(y_{ki}|T_i \geqslant t_k; \mathbf{Z}_{ki}), \quad k=1,\ldots,K;\ i=1,\ldots,n$$

Then mimicking the generalized estimating equations with independence working assumption [6], we propose the following estimating function for $\boldsymbol{\beta}$ of model (6):

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} I(X_i \geqslant t_k) h(\mathbf{Z}_{ki}; \boldsymbol{\beta})\{y_{ki} - g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki} \tag{8}$$

where h is a given scalar function. The choices of h are discussed in McCullagh and Nelder [7] and Liang and Zeger [6] for $K=1$ and $K>1$, respectively. For the aforementioned linear and multiplicative models, it is reasonable to set $h=1$.

By the law of conditional expectations

$$\begin{aligned} E\{\mathbf{U}(\boldsymbol{\beta})\} &= E\left[\sum_{i=1}^{n}\sum_{k=1}^{K} I(X_i \geqslant t_k) h(\mathbf{Z}_{ki}; \boldsymbol{\beta})\{E(y_{ki}|X_i \geqslant t_k; \mathbf{Z}_{ki}) - g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}\right] \\ &= E\left[\sum_{i=1}^{n}\sum_{k=1}^{K} I(X_i \geqslant t_k) h(\mathbf{Z}_{ki}; \boldsymbol{\beta})\{E(y_{ki}|T_i \geqslant t_k; \mathbf{Z}_{ki}) - g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}\right] = 0 \end{aligned}$$

Thus, $\mathbf{U}(\boldsymbol{\beta})$ is an unbiased estimating function for $\boldsymbol{\beta}$. Denote the solution to $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$ by $\hat{\boldsymbol{\beta}}$. By standard asymptotic arguments [6], $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically zero-mean normal with a covariance matrix consistently estimated by

$$n\{\partial\mathbf{U}(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}'\}^{-1}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} I(X_i \geqslant t_k) h(\mathbf{Z}_{ki}; \hat{\boldsymbol{\beta}})\{y_{ki} - g(\hat{\boldsymbol{\beta}}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}\right]^{\otimes 2}\{\partial\mathbf{U}(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}'\}^{-1}$$

Here and in the following, we adopt the notation $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$ and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$.

Inference under model (1) is more delicate. If all the y_{ki}'s were known, then the generalized estimating equation (with independence working assumption) for $\boldsymbol{\beta}$ of (1) would take the form

$$\sum_{i=1}^{n}\sum_{k=1}^{K} h(\mathbf{Z}_{ki}; \boldsymbol{\beta})\{y_{ki} - g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki} = \mathbf{0}$$

In the presence of censoring, not all y_{ki} are observable, so that the above equation ought to be modified. Let $T_{ki}^* = T_i \wedge t_k$, and $\delta_{ki}^* = I(C_i \geqslant T_{ki}^*)$. Clearly, y_{ki} is known if and only if $\delta_{ki}^* = 1$.

Define $\mathbf{F}_i(t)=\{I(T_i\leqslant t), Y_i(t), \mathbf{L}_i(t)\}$ and $G(t|\bar{\mathbf{F}}_i)=\Pr\{C_i\geqslant t|\bar{\mathbf{F}}_i(T_i)\}$, where $\mathbf{L}_i(.)$ denotes all the measured covariate processes and $\bar{\mathbf{H}}(t)=\{\mathbf{H}(s)\colon s\leqslant t\}$ for any process $\mathbf{H}(.)$. We then propose the following estimating function:

$$\mathbf{U}(\boldsymbol{\beta})=\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^{*}}{\hat{G}(T_{ki}^{*}|\bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki} \tag{9}$$

where $\hat{G}(t|\bar{\mathbf{F}})$ is a consistent estimator of $G(t|\bar{\mathbf{F}})$. For notational simplicity, the resulting estimator is also denoted by $\hat{\boldsymbol{\beta}}$.

The proposed estimating function includes only the non-censored y_{ki}'s, but their contributions are weighted by their (estimated) inverse probabilities of inclusion. This is commonly referred to as inverse probability of censoring weighting (IPCW) and is reminiscent of the Horvitz–Thompson estimator [8]. The IPCW technique was previously used by Koul *et al.* [9], Robins and Rotnitzky [10], Lin and Ying [11] and Zhao and Tsiatis [12] in different contexts. The estimating function for model (2) developed by Lin [3] is a special case of (9).

If censoring occurs in a completely random fashion, we may set $\hat{G}(\cdot|\bar{\mathbf{F}})$ to be the Kaplan–Meier estimator $\hat{G}(.)$ for the common survival function of C_i. Otherwise, we assume that the hazard function corresponding to $G(t|\bar{\mathbf{F}}_i)$ satisfies the proportional hazards model [4]

$$\lambda(t|\bar{\mathbf{F}}_i)=\lambda_0(t)\mathrm{e}^{\gamma'\mathbf{W}_i(t)}, \quad i=1,\ldots,n \tag{10}$$

where $\mathbf{W}_i(t)$ is a vector of known functions of $\mathbf{F}_i(t)$, γ is a vector of unknown regression parameters and $\lambda_0(.)$ an arbitrary baseline hazard function. We then set $\hat{G}(t|\bar{\mathbf{F}}_i)$ to be the Breslow estimator [13]

$$\hat{G}(t|\mathbf{W}_i)\equiv\exp\left\{-\sum_{j=1}^{n}\frac{\bar{\delta}_j I(X_j<t)\mathrm{e}^{\hat{\gamma}'\mathbf{W}_i(X_j)}}{S^{(0)}(X_j;\hat{\gamma})}\right\}$$

where $\bar{\delta}_i=1-\delta_i$, $\hat{\gamma}$ is the maximum partial likelihood estimator of γ, and

$$\mathbf{S}^{(\rho)}(t;\gamma)=\sum_{i=1}^{n}I(X_i\geqslant t)\mathrm{e}^{\gamma'\mathbf{W}_i(t)}\mathbf{W}_i^{\otimes\rho}(t), \quad \rho=0,1,2$$

In the Appendix, we show that $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically zero-mean normal with a covariance matrix consistently estimated by $n\{\partial\mathbf{U}(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}'\}^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i^{\otimes 2}\{\partial\mathbf{U}(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}'\}^{-1}$, where the form of $\hat{\boldsymbol{\xi}}_i$ depends on how $\hat{G}$ is calculated. If $\hat{G}$ is the Kaplan–Meier estimator, then

$$\hat{\boldsymbol{\xi}}_i=\sum_{k=1}^{K}\frac{\delta_{ki}^{*}h(\mathbf{Z}_{ki};\hat{\boldsymbol{\beta}})\{y_{ki}-g(\hat{\boldsymbol{\beta}}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}}{\hat{G}(T_{ki}^{*})}+\bar{\delta}_i\mathbf{Q}(X_i)-\sum_{j=1}^{n}\frac{\bar{\delta}_j I(X_j\leqslant X_i)\mathbf{Q}(X_j)}{\sum_{l=1}^{n}I(X_l\geqslant X_j)}$$

where

$$\mathbf{Q}(t)=\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{I(T_{ki}^{*}>t)\delta_{ki}^{*}h(\mathbf{Z}_{ki};\hat{\boldsymbol{\beta}})\{y_{ki}-g(\hat{\boldsymbol{\beta}}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}}{\hat{G}(T_{ki}^{*})}\Bigg/\sum_{j=1}^{n}I(X_j\geqslant t)$$

If $\hat{G}$ is the Breslow estimator, then

$$\hat{\boldsymbol{\xi}}_i=\sum_{k=1}^{K}\frac{\delta_{ki}^{*}h(\mathbf{Z}_{ki};\hat{\boldsymbol{\beta}})\{y_{ki}-g(\hat{\boldsymbol{\beta}}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}}{\hat{G}(T_{ki}^{*}|\mathbf{W}_i)}+\bar{\delta}_i\mathbf{D}_i(X_i)-\sum_{j=1}^{n}\frac{\bar{\delta}_j I(X_j\leqslant X_i)\mathrm{e}^{\hat{\gamma}'\mathbf{W}_i(X_j)}\mathbf{D}_i(X_j)}{S^{(0)}(X_j;\hat{\gamma})}$$

where $\mathbf{D}_i(t)=\mathbf{Q}(t)+\mathbf{B}\mathbf{\Omega}^{-1}\{\mathbf{W}_i(t)-\mathbf{S}^{(1)}(t;\hat{\gamma})/S^{(0)}(t;\hat{\gamma})\}$

$$\mathbf{Q}(t)=\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{I(T_{ki}^{*}>t)\mathrm{e}^{\hat{\gamma}'\mathbf{W}_i(t)}\delta_{ki}^{*}h(\mathbf{Z}_{ki};\hat{\boldsymbol{\beta}})\{y_{ki}-g(\hat{\boldsymbol{\beta}}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}}{\hat{G}(T_{ki}^{*}|\mathbf{W}_i)S^{(0)}(t;\hat{\gamma})}$$

$$\mathbf{B}=\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^{*}h(\mathbf{Z}_{ki};\hat{\boldsymbol{\beta}})\{y_{ki}-g(\hat{\boldsymbol{\beta}}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}\mathbf{R}'(T_{ki}^{*};\mathbf{W}_i)}{\hat{G}(T_{ki}^{*}|\mathbf{W}_i)}$$

$$\mathbf{R}(t;\mathbf{W})=\sum_{i=1}^{n}\bar{\delta}_i I(X_i<t)\mathrm{e}^{\hat{\gamma}'\mathbf{W}(X_i)}\left\{\mathbf{W}(X_i)-\frac{\mathbf{S}^{(1)}(X_i;\hat{\gamma})}{S^{(0)}(X_i;\hat{\gamma})}\right\}\Big/S^{(0)}(X_i;\hat{\gamma})$$

and

$$\hat{\mathbf{\Omega}}=\sum_{i=1}^{n}\bar{\delta}_i\left\{\frac{\mathbf{S}^{(2)}(X_i;\hat{\gamma})}{S^{(0)}(X_i;\hat{\gamma})}-\frac{\mathbf{S}^{(1)}(X_i;\hat{\gamma})^{\otimes 2}}{S^{(0)}(X_i;\hat{\gamma})^2}\right\}$$

If $\delta_i=1$, then both Y_i and $I(t_{k-1}<T_i\leqslant t_k)$ are known. Thus, we propose to estimate $\boldsymbol{\beta}$ of model (7) by the following analogue of (9):

$$\mathbf{U}(\boldsymbol{\beta})=\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_i I(t_{k-1}<T_i\leqslant t_k)}{\hat{G}(T_i|\bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{Y_i-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki} \qquad (11)$$

The resulting estimator, again denoted by $\hat{\boldsymbol{\beta}}$, possesses the aforementioned asymptotic properties for the root of (9), but δ_{ki}^{*} and y_{ki} ($k=1,\ldots,K$; $i=1,\ldots,n$) are replaced by $\delta_i I(t_{k-1}<T_i\leqslant t_k)$ and Y_i throughout.

As evident from (11), the estimation of model (7) with interval-specific parameters requires a few observed deaths in each time interval. Thus, the intervals cannot be too fine for small samples with heavy censoring. If the study period is not partitioned at all, then model (7) becomes

$$E(Y_i|T_i\leqslant\tau;\mathbf{Z}_i)=g(\boldsymbol{\beta}'\mathbf{Z}_i) \qquad (12)$$

When intervals are broad, especially for model (12), it may be necessary to include some functions of survival time in the covariate vector since medical cost tends to be highly correlated with survival time. The relationship between survival time and cost is often complex. Thus, for large studies such as the linked SEER-Medicare database, it is desirable to use model (7) with fine intervals so that the effects of survival time on cost needs not be parameterized.

3. NUMERICAL STUDIES

Extensive simulation studies were carried out to evaluate the finite-sample properties of the inference procedures under models (1), (6) and (7). In general, these three models do not hold with the same set of regression parameters. For simplicity of description, we present here the simulation results for the situations in which the covariate effects are the same under the three models.

Survival and censoring times were generated from the exponential distribution with mean m and the uniform $(0,c)$ distribution, respectively. We set $\tau=10$, by which point all survival times and cost accumulation processes are censored. The combinations of $(m,c)=(5,40)$, $(5,20)$, $(10, 40)$ and $(10,20)$ yield approximately 20, 30, 40 and 50 per cent censored survival times.

The entire study period $(0,10]$ was divided into ten equally spaced intervals. We set

$$\begin{aligned} y_{ki} = {} & [I(k=1)u_i^d + I(T_i > t_k)(\eta_i + u_{ki}) \\ & + I(t_{k-1} < T_i \leqslant t_k)\{(\eta_i + u_{ki})(T_i - t_{k-1}) + u_i^f\}]\mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_i}, \quad k=1,\ldots,K;\ i=1,\ldots,n \end{aligned}$$

where η_i, u_{ki}, u_i^d and u_i^f are independent random variables with the uniform (0,1) distribution for η_i and u_{ki} and uniform (0,5) and (0,10) distributions for u_i^d and u_i^f, respectively. This scheme creates J-shaped time patterns; there is some basic cost for each time interval in which the subject is alive, in addition, there is a relatively high diagnostic cost for the first interval and an even higher final cost for the interval in which the subject dies. For the same subject, the basic costs in different intervals share a common random effect and are thus positively correlated.

It is easy to see that the cost data so generated satisfy models (1), (6) and (7) with the same covariate effects. Specifically

$$E(y_{ki}|\mathbf{Z}_i) = \mu_k \mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_i}, \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{13}$$

$$E(y_{ki}|T_i > t_k; \mathbf{Z}_i) = \mu_k \mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_i}, \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{14}$$

$$E(Y_i|t_{k-1} < T_i \leqslant t_k; \mathbf{Z}_i) = \mu_k \mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_i}, \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{15}$$

Although the μ_k's are different among these three models, $\boldsymbol{\beta}$ is the same.

In our main studies, Z was set to be a treatment indicator with $n/2$ subjects in each of the two groups, and β was set to 1. We chose $n=100$, 200 and 500. For $n=100$ and 200, it is not always possible to fit model (15) because some intervals contain no death. Thus, we considered the following special case of model (12):

$$E(Y_i|T_i \leqslant \tau; Z_i, T_i) = \mathrm{e}^{\alpha+\beta Z_i+\gamma T_i} \tag{16}$$

which also holds for the generated cost data. We fit model (16) under $n=100$, 200 and 500, and fit model (15) only under $n=500$. We set $h(Z;\beta)=1$ for all estimators.

The results from the main studies are summarized in Table I. The parameter estimators are virtually unbiased in all cases. The standard error estimators reflect well the true variabilities of the parameter estimators and the associated Wald-type confidence intervals have proper coverage probabilities, at least for $n \geqslant 200$. The use of model (16) rather than (15) improves the efficiency of the parameter estimation and the accuracy of the asymptotic approximation, though the efficiency gain is minimal. Under models (13) and (14), the variances of the estimators appear to decrease as m or c increases; under models (15) and (16), the variances tend to decrease as the amount of censoring decreases.

Table I. Simulation results for the estimation of β under models (13)–(16).

m	c	n	Model (13)				Model (14)				Model (15) or (16)			
			Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
5	40	100	−0.003	0.82	0.81	0.944	−0.005	0.72	0.70	0.939	0.003	0.74	0.73	0.941
		200	0.003	0.58	0.58	0.948	0.004	0.51	0.50	0.948	0.003	0.52	0.52	0.947
		500	0.007	0.37	0.37	0.947	−0.001	0.32	0.32	0.947	−0.001	0.33	0.33	0.947
											−0.001	0.34	0.33	0.945
5	20	100	−0.014	0.90	0.89	0.941	−0.006	0.74	0.72	0.940	0.003	0.80	0.78	0.940
		200	0.001	0.64	0.63	0.945	0.003	0.52	0.51	0.945	0.003	0.56	0.55	0.945
		500	0.004	0.40	0.40	0.948	−0.002	0.33	0.33	0.946	−0.001	0.35	0.35	0.948
											−0.041	0.38	0.37	0.943
10	40	100	−0.004	0.75	0.74	0.943	0.000	0.65	0.63	0.940	0.004	0.85	0.82	0.940
		200	−0.003	0.52	0.52	0.950	0.003	0.46	0.45	0.944	−0.002	0.59	0.58	0.941
		500	0.003	0.33	0.33	0.949	−0.000	0.29	0.29	0.947	−0.003	0.37	0.37	0.947
											−0.003	0.41	0.40	0.941
10	20	100	−0.012	0.81	0.80	0.944	−0.001	0.67	0.65	0.941	0.004	0.96	0.90	0.930
		200	−0.002	0.57	0.57	0.946	0.003	0.47	0.46	0.942	0.001	0.65	0.64	0.943
		500	0.004	0.36	0.36	0.948	−0.001	0.30	0.30	0.947	−0.002	0.40	0.40	0.946
											−0.004	0.42	0.41	0.940

Bias is the mean of $\hat{\beta}$ minus β; SE is the standard error of $\hat{\beta}$; SEE is the mean of the standard error estimator for $\hat{\beta}$; CP is the coverage probability of the 95 per cent confidence interval for β. Bias, SE and SEE are multiplied by 10. Under the heading 'Model (15) or (16)', all the results pertain to model (16) except for the last row of each block, which pertains to model (15). Each entry is based on 10000 simulation samples.

Additional studies revealed that the results were virtually unchanged when β was set to other values. Furthermore, the asymptotic approximations continued to be accurate when covariates, survival times, censoring times and costs were generated from other distributions.

4. SEER-MEDICARE DATABASE

We now apply the proposed methods to a subset of data from the aforementioned SEER-Medicare database. This subset consists of all the 3550 Medicare beneficiaries over the age of 65 who were diagnosed with epithelial ovarian cancer during the years 1984–1989. Out of these subjects, 540, 836 and 2174 were diagnosed with local, regional and distant stages, respectively. It is of public-health importance to understand how the clinical stage at diagnosis affects the subsequent survival experience and medical cost.

As mentioned in Section 1, the data on survival time and monthly medical expenditures were recorded from 1984 to 1990. The subjects who were still alive at the end of 1990 were censored. Although the follow-up was terminated at the same time point, the actual censoring times, as measured from the times of diagnosis, varied substantially among the subjects because the times of diagnosis spanned a period of 6 years. There was no voluntary loss to follow-up in this study, so that censoring, which was solely caused by limited study

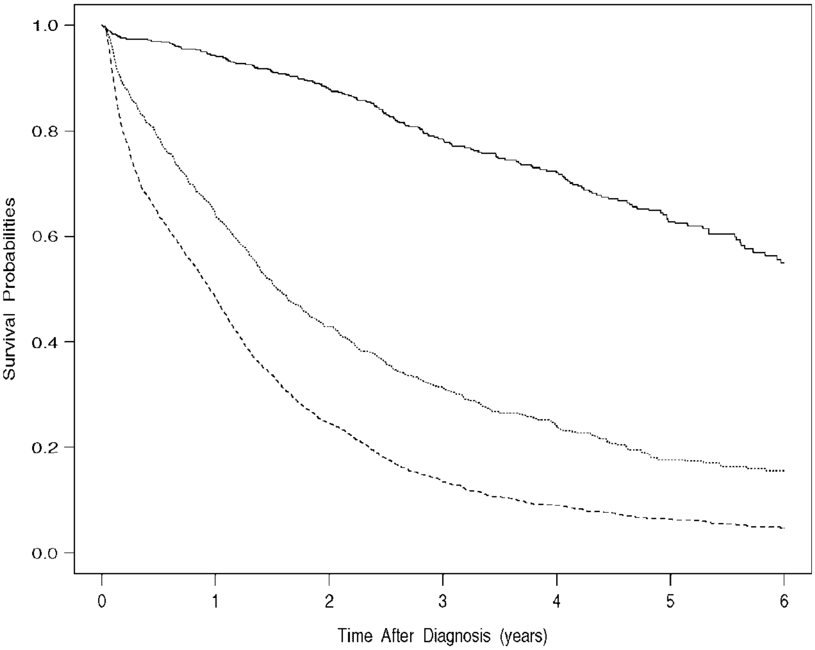

Figure 3. Kaplan–Meier estimates of the survival curves for ovarian cancer patients: — local stage; ··· regional stage; ––– distant stage.

duration, can be regarded as completely random. Thus, the proposed methods with $\hat{G}$ as the Kaplan–Meier estimator can be used.

Although the follow-up extended over a period of almost 7 years, the data for the last year are very sparse, especially for the regional- and distant-stage patients, most of whom did not survive to the 7th year. Thus, we confine our attention to the first 6 years after diagnosis. With the cost data recorded in monthly intervals, there are a total of 72 intervals over the 6-year period.

To assess the differences among the three stages, we let $\mathbf{Z}_i$ consist of two indicator covariates with local stage as the reference group. If we allow both the intercepts and effects of $\mathbf{Z}_i$ to be interval-specific, then the model will be saturated with 216 parameters and the estimates so obtained will be basically equivalent to the non-parametric (one-sample) estimates over the 72 intervals separated by the three stages. The question is whether the data can be characterized by more parsimonious models.

Figure 3 displays the Kaplan–Meier estimates of the survival curves for the three stages. The patients with less advanced disease tend to live longer. Under the proportional hazards model, the hazard ratios for regional and distant stages versus local stage are estimated at 3.86 and 6.56, respectively, with very small standard error estimates. The 6-year survival probabilities are approximately 55, 15 and 5 per cent for the local, regional and distant stages, respectively.

Figure 4 presents the non-parametric estimates of the mean cumulative costs for the three stages, which are obtained by calculating $\{\sum \delta^*_{ki} y_{ki}/\hat{G}(T^*_{ki})\}/\{\sum \delta^*_{ki}/\hat{G}(T^*_{ki})\}$ $(k=1,\ldots,72)$ within each of the three groups and then accumulating them over the 72 intervals. The mean

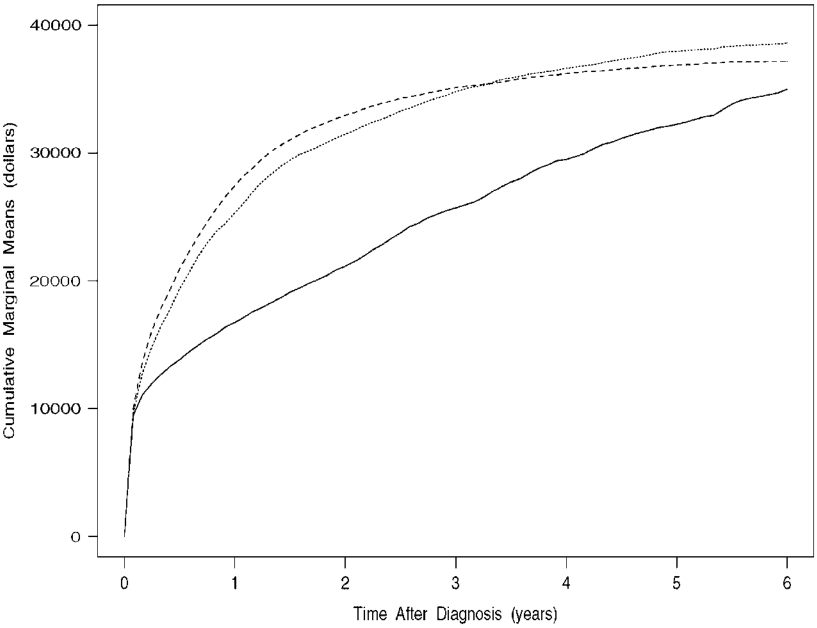

Figure 4. Non-parametric estimates of the cumulative marginal means of costs for ovarian cancer patients: — local stage; ··· regional stage; −−− distant stage.

Table II. Regression estimates for the proportionate differences of the regional- and distant-stage patients from the local-stage patients in medical costs.

Model	Stage	β				e^{β}	
		Est	SE	Est/SE	95% CI	Est	95% CI
(13)	Regional	0.106	0.053	2.02	(0.003, 0.209)	1.11	(1.01, 1.23)
	Distant	0.058	0.047	1.25	(−0.033, 0.150)	1.06	(0.97, 1.16)
(14)	Regional	0.554	0.046	11.95	(0.463, 0.645)	1.74	(1.59, 1.91)
	Distant	0.758	0.040	18.85	(0.680, 0.837)	2.13	(1.97, 2.31)
(15)	Regional	0.138	0.068	2.06	(0.006, 0.271)	1.15	(1.01, 1.31)
	Distant	0.212	0.063	3.34	(0.088, 0.336)	1.24	(1.09, 1.40)

Local stage is the reference group. Est, SE and CI stand for estimate, (estimated) standard error and confidence interval, respectively. The weights $h(\mathbf{Z}_{ki}, \boldsymbol{\beta})$ are set to 1 for all three models.

costs for the regional and distant stages are higher than those of the local stage for the first two years but lower afterwards. The effects of covariates are not constant over time, at least not on an additive or multiplicative scale.

The top panel of Table II provides the estimates of the common covariate effects under model (13). These results are not terribly interesting since the model does not hold. The

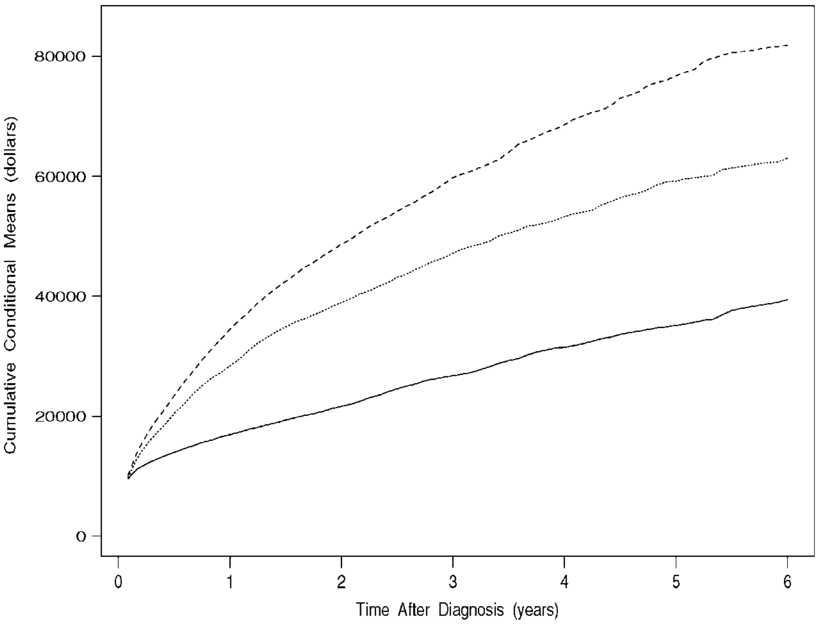

Figure 5. Non-parametric estimates of the cumulative conditional means of costs among survivors for ovarian cancer patients: — local stage; ··· regional stage; ––– distant stage.

model misspecification encountered here is similar to the violation of the proportional hazards assumption for the Cox model. What is being estimated is an average of the (multiplicative) covariate effects over time. Such an average does not adequately represent the actual covariate effects when the effects are not in the same direction over time, as is clearly the case here.

As evident from Figure 3, the local-stage patients tend to live much longer than the regional- and distant-stage patients, and thus have more opportunities to incur medical costs. This fact may well explain the phenomenon seen in Figure 4. Thus, it is useful to compare the costs among patients who have similar survival experiences, as argued in Sections 1 and 2.

Figure 5 displays the accumulation over time of the non-parametric estimates for the conditional means of the costs among the survivors, $\sum I(X_i \geqslant t_k)y_{ki}/\sum I(X_i \geqslant t_k)$ $(k=1,\ldots,72)$, separated by three disease stages. The fact that the three curves appear to be proportionate over time suggests that model (14) provides a reasonable description of the data. The estimates from this model are shown in the middle panel of Table II. The conditional means for the regional stage are approximately 75 per cent higher than those of the local stage, and the conditional means for the distant stage are more than double those of the local stage.

Figure 6 shows the accumulation of the non-parametric estimates for the conditional means among the deaths, $\{\sum \delta_i I(t_{k-1} < T_i \leqslant t_k)Y_i/\hat{G}(T_i)\}/\{\sum \delta_i I(t_{k-1} < T_i \leqslant t_k)/\hat{G}(T_i)\}$ $(k=1,\ldots,72)$. The three curves seem roughly proportionate; however, the trends are not as clear as in Figure 5 (for example, the differences between the regional stage and local stage appear to be constant rather than increasing as time varies from year 3 to year 6). Thus, model (15) is reasonable, though not entirely satisfactory. The results from this model are given in the bottom panel of Table II. Based on the point estimates, the conditional means for the regional

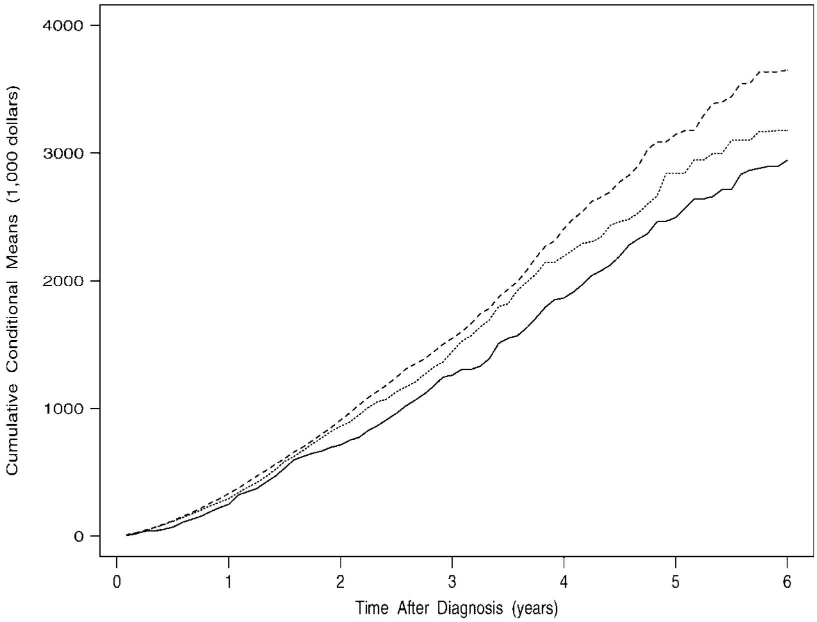

Figure 6. Non-parametric estimates of the cumulative conditional means of costs among deaths for ovarian cancer patients: — local stage; ··· regional stage; ––– distant stage.

and distant stages are, respectively, 15 per cent and 24 per cent higher than those of the local stage. The level of statistical significance is lower under model (15) than under (14).

The above analysis indicates that the local-stage survivors tend to incur costs at much slower rates than the regional-stage survivors and, to an even greater extent, the distant-stage survivors. Among those who die in the same month, the total post-diagnosis costs for the local-stage subjects tend to be slightly lower than those of the regional- and distant-stage subjects. Because the subjects with less aggressive disease live longer, the long-term costs for the local-stage subjects may well be higher than those of the regional- and distant-stage subjects.

The results provided by conditional models (14) and (15) reveal that the interactions between time and stage effects in the marginal mean, as evident in Figure 4, is attributed primarily to the fact that the survival rates are different among the three disease stages. The local-stage patients tend to accumulate costs at slower rates than the regional- and distant-stage patients both when they are alive and around the time of death. Thus, the cumulative marginal mean function for the local stage increases over time less rapidly than those of the regional and distant stages, as shown in Figure 4. After two years, however, the curves begin to converge because the local-stage patients tend to live longer so that their long-term costs may catch up with or even surpass those of the regional- and distant-stage patients.

To further illustrate the proposed methodology, we add to models (13)–(15) a continuous covariate, the time of first diagnosis. The results of the analysis are summarized in Table III. There appears to be a negative time trend; the patients who are diagnosed later in calendar time tend to have lower subsequent medical costs. The changes are fairly small, although

Table III. Regression estimates for the proportionate effects of clinical stage and time of first diagnosis on medical costs.

Model	Parameter	β				e^{β}	
		Est	SE	Est/SE	95% CI	Est	95% CI
(13)	Regional	0.101	0.052	1.94	$(-0.001, 0.203)$	1.11	(1.00, 1.22)
	Distant	0.090	0.046	1.98	(0.001, 0.180)	1.09	(1.00, 1.20)
	Time	-0.060	0.008	-7.65	$(-0.075, -0.044)$	0.94	(0.93, 0.96)
(14)	Regional	0.540	0.047	11.59	(0.448, 0.631)	1.71	(1.57, 1.88)
	Distant	0.759	0.041	18.69	(0.680, 0.839)	2.14	(1.97, 2.31)
	Time	-0.021	0.007	-3.05	$(-0.035, -0.008)$	0.98	(0.97, 0.99)
(15)	Regional	0.132	0.068	1.95	$(-0.001, 0.264)$	1.14	(1.00, 1.30)
	Distant	0.221	0.063	3.48	(0.096, 0.345)	1.25	(1.10, 1.41)
	Time	-0.024	0.008	-2.94	$(-0.039, -0.008)$	0.98	(0.96, 0.99)

Regional and Distant compare the regional and distant stages with the local stage. Time refers to the time (in years) of first diagnosis. Est, SE and CI stand for estimate, (estimated) standard error and confidence interval, respectively. The weights $h(\mathbf{Z}_{ki}, \boldsymbol{\beta})$ are set to 1 for all three models.

highly significant. The inclusion of the time variable has little effects on the estimates for the differences among the three clinical stages.

5. DISCUSSION

In this paper we present three classes of models which address different aspects of the associations between covariates and cost accumulation. The marginal mean is highly important from the view of public health as it reflects the actual costs. The conditional means are more informative about the underlying mechanism of cost accumulation, but do not readily translate into ultimate costs. The effects of covariates on the marginal and conditional means can be quite different especially when the covariates have substantial effects on survival time. The pattern-mixture models separate the covariate effects on survival time from the covariate effects on the rate of cost accumulation, whereas the marginal models pertain to the ultimate costs given the actual survival distributions. As evident from the previous section, the application of the three types of models to the same data set can provide more insights into the effects of covariates on cost accumulation than the use of a single model.

The SEER-Medicare data was previously analysed by Etzioni *et al.* [14] and Lin [3]. Specifically, Etzioni *et al.* used the method of Lin *et al.* [2] to estimate the marginal mean costs for the three clinical stages, while Lin [3] used the linear regression to estimate the differences in the mean 5-year cost. The method of Lin *et al.* [2] requires that the censoring time be discrete. As discussed in Sections 1 and 2, the linear model employed by Lin [3] has severe limitations; neither Etzioni *et al.* [14] nor Lin [3] were able to examine the conditional distributions of cost accumulation given specific survival patterns.

The models considered in this paper, although flexible and versatile, are by no means exhaustive. One may model the marginal means of the Y_{ki}'s rather than the y_{ki}'s through

$$E(Y_{ki}|\mathbf{Z}_{ki}) = g(\boldsymbol{\beta}'\mathbf{Z}_{ki}), \quad k=1,\ldots,K;\ i=1,\ldots,n \tag{17}$$

As evident from the discussion in Section 2.2, (17) is equivalent to (1) only under special circumstances. Instead of conditioning on survival to the end of the interval, as in model (6), one may condition on survival to the beginning of the interval and consider models of the form

$$E(y_{ki}|T_i > t_{k-1}; \mathbf{Z}_{ki}) = g(\boldsymbol{\beta}'\mathbf{Z}_{ki}), \quad k=1,\ldots,K;\ i=1,\ldots,n$$

A more elaborate version of (7) would be

$$E(y_{li}|t_{k-1} < T_i \leqslant t_k; \mathbf{Z}_{kli}) = g(\boldsymbol{\beta}'\mathbf{Z}_{kli}), \quad k=1,\ldots,K;\ l=1,\ldots,k;\ i=1,\ldots,n$$

under which covariates are allowed to have different effects for the same time interval among subjects who die in different intervals. Inference procedures for these new models as well as other types of marginal models can be developed along the lines of Section 2.3. Other quantities which one might model, and comments on their interpretability, are given by Cook and Lawless [15] in the context of recurrent events.

The proposed estimators for models (1) and (7) involve inverse probability weighting. This kind of weighting generally does not lead to efficient estimators. In some cases, one can obtain efficient estimators if the weights are estimated in the right way; see Robins and Rotnitzky [10]. The special features of the cost data, including the presence of death and the dependence of responses, means that efficient estimating functions would take much more complicated forms than the ones given here. Bang and Tsiatis [16] studied efficient estimation of the marginal mean in the one-sample case. The generalization of their results to the regression setting is not straightforward, but is certainly worth pursuing.

In Section 4, we used the non-parametric estimates of the cumulative means to assess the adequacy of the regression models. This approach is generally applicable, although continuous covariates would need to be discretized for the purposes of model checking. The pragmatic visual diagnostic could be improved by adding error bounds to the estimated curves. It should be noted that the cumulative conditional means shown in Figures 5 and 6 are not interpretable in terms of anything observable. Formal model checking can be performed by testing extra parameters, such as covariates × time interactions, in embedded models [7] or by considering certain aggregates of residuals [17].

Lin [18] proposed the following continuous-time model for the marginal distribution of the accumulation process:

$$E\{Y_i(t)|\mathbf{Z}_i\} = \mu_0(t)\mathrm{e}^{\boldsymbol{\beta}'\mathbf{Z}_i}, \quad i=1,\ldots,n \tag{18}$$

where $\mu_0(.)$ is an arbitrary positive function, and developed simple methods for making inference about $\boldsymbol{\beta}$. This model is somewhat restrictive in that it imposes common proportionate covariate effects over time, and thus would not be appropriate for the SEER-Medicare data. Furthermore, the inference for $\boldsymbol{\beta}$ requires that censoring arise in a completely random fashion. Model (18) implies models (4) and (5) with $\mu_k = \mu_0(t_k) - \mu_0(t_{k-1})$. Unlike models (4) and (5), the adequacy of model (18) cannot be checked empirically unless the entire sample paths of $\{Y_i(t); 0 < t \leqslant C_i\}$ $(i=1,\ldots,n)$ are observed.

Semi-parametric inference with censored medical cost data is possible only under the following assumption on censoring:

$$\lambda\{t|\bar{\mathbf{F}}_i(T_i)\}=\lambda\{t|\bar{\mathbf{F}}_i(t)\}, \quad t\leqslant T_i;\ i=1,\ldots,n$$

This assumption allows the probability of censoring over $[t, t+\mathrm{d}t)$ to depend on the history up to t of the survival and cost accumulation processes as well as any measured covariate processes. To estimate models (1) and (7), we formulate $\lambda\{t|\bar{\mathbf{F}}_i(t)\}$ through (10), although any other parametric or semi-parametric models may also be used. (If covariates are all discrete, then non-parametric models may be used.) The adequacy of model (10) can be assessed via existing goodness-of-fit methods for the proportional hazards model [19]. To estimate model (6), we set $\lambda\{t|\bar{\mathbf{F}}_i(t)\}=\lambda\{t|\bar{\mathbf{Z}}_i(t)\}$; under this restrictive assumption, the form of $\lambda\{t|\bar{\mathbf{Z}}_i(t)\}$ need not be specified at all. If censoring depends on other covariates besides $\mathbf{Z}_i$, then one needs to formulate $\lambda\{t|\bar{\mathbf{F}}_i(t)\}$ and then apply the IPCW technique to model (6). It is necessary to use the IPCW technique for models (1) and (7) even if censoring is purely random.

This paper extends the prior work of Lin [3, 18] in several important directions. For modelling the marginal means, model (1) is much more flexible and versatile than model (2); the existing inference procedures for model (18) require censoring to be completely random and may not have good efficiencies. Models (6) and (7) are brand new, and allow one to separate the effects of covariates on the survival time from the effects of covariates on the rate of cost accumulation.

The estimation of medical cost studied in this paper is an important component in the cost-effectiveness analysis. A number of authors have discussed how to measure and estimate cost-effectiveness [20–23]. By combining their ideas with those presented in this paper, it is possible to perform the cost-effectiveness analysis with covariate adjustments based on censored data. The details will be communicated in a separate paper.

APPENDIX: PROOFS OF ASYMPTOTIC RESULTS FOR MODEL (1)

Let us make the decomposition $\mathbf{U}(\boldsymbol{\beta})=\mathbf{U}_1(\boldsymbol{\beta})+\mathbf{U}_2(\boldsymbol{\beta})$, where

$$\begin{aligned}\mathbf{U}_1(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^*}{G(T_{ki}^*|\bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}\\ \mathbf{U}_2(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{G(T_{ki}^*|\bar{\mathbf{F}}_i)-\hat{G}(T_{ki}^*|\bar{\mathbf{F}}_i)}{G(T_{ki}^*|\bar{\mathbf{F}}_i)\hat{G}(T_{ki}^*|\bar{\mathbf{F}}_i)}\delta_{ki}^*h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}\end{aligned} \tag{A1}$$

The fact that $E(\delta_{ki}^*|\bar{\mathbf{F}}_i)=G(T_{ki}^*|\bar{\mathbf{F}}_i)$ implies that $\mathbf{U}_1(\boldsymbol{\beta})$ is a sum of n independent zero-mean random vectors under model (1). Thus, by the law of large numbers, $n^{-1}\mathbf{U}_1(\boldsymbol{\beta})$ converges in probability to $\mathbf{0}$. In addition, it follows from the consistency of the Kaplan–Meier and Breslow estimators that $n^{-1}\mathbf{U}_2(\boldsymbol{\beta})$ also converges in probability to $\mathbf{0}$. Suppose that $-n^{-1}\partial\mathbf{U}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$ is positive definite, at least for large n. It then follows from convex analysis that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.

Under model (10)

$$\frac{n^{1/2}\{G(t|\mathbf{W})-\hat{G}(t|\mathbf{W})\}}{G(t|\mathbf{W})}=n^{-1/2}\sum_{i=1}^{n}\int_{0}^{t}\frac{\mathrm{e}^{\gamma'\mathbf{W}(x)}\,\mathrm{d}M_i(x)}{s^{(0)}(x)}+\mathbf{r}'(t;\mathbf{W})\mathbf{\Omega}^{-1}$$
$$\times n^{-1/2}\sum_{i=1}^{n}\int_{0}^{\infty}\{\mathbf{W}_i(x)-\bar{\mathbf{w}}(x)\}\,\mathrm{d}M_i(x)+o_p(1) \qquad \text{(A2)}$$

where $\mathbf{s}^{(\rho)}(t)=\lim_{n\to\infty}n^{-1}\mathbf{S}^{(\rho)}(t;\gamma)$ $(\rho=0,1,2)$, $\bar{\mathbf{w}}(t)=\mathbf{s}^{(1)}(t)/s^{(0)}(t)$

$$\mathbf{r}(t;\mathbf{W})=\int_{0}^{t}\mathrm{e}^{\gamma'\mathbf{W}(x)}\{\mathbf{W}(x)-\bar{\mathbf{w}}(x)\}\lambda_0(x)\,\mathrm{d}x$$

$$\mathbf{\Omega}=\int_{0}^{\infty}\{\mathbf{s}^{(2)}(t)/s^{(0)}(t)-\bar{\mathbf{w}}^{\otimes 2}(t)\}s^{(0)}(t)\lambda_0(t)\,\mathrm{d}t$$

and $M_i(t)=\bar{\delta}_i I(X_i\leqslant t)-\int_0^t I(X_i\geqslant x)\mathrm{e}^{\gamma'\mathbf{W}_i(x)}\lambda_0(x)\,\mathrm{d}x$ [24]. In view of (A2)

$$n^{-1/2}\mathbf{U}_2(\boldsymbol{\beta})=n^{-1/2}\sum_{i=1}^{n}\int_{0}^{\infty}\tilde{\mathbf{Q}}(t)\,\mathrm{d}M_i(t)+\tilde{\mathbf{B}}\mathbf{\Omega}^{-1}$$
$$\times n^{-1/2}\sum_{i=1}^{n}\int_{0}^{\infty}\{\mathbf{W}_i(t)-\bar{\mathbf{w}}(t)\}\,\mathrm{d}M_i(t)+o_p(1)$$

where

$$\tilde{\mathbf{Q}}(t)=n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{I(T_{ki}^{*}>t)\mathrm{e}^{\gamma'\mathbf{W}_i(t)}\delta_{ki}^{*}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}}{\hat{G}(T_{ki}^{*}|\mathbf{W}_i)s^{(0)}(t)}$$

and

$$\tilde{\mathbf{B}}=n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^{*}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}\mathbf{r}'(T_{ki}^{*};\mathbf{W}_i)}{\hat{G}(T_{ki}^{*}|\mathbf{W}_i)}$$

By the law of large numbers and the consistency of the Breslow estimator, $\tilde{\mathbf{Q}}(t)$ and $\tilde{\mathbf{B}}$ converge in probability to deterministic limits, say $\mathbf{q}(t)$ and $\mathbf{b}$. Thus

$$n^{-1/2}\mathbf{U}_2(\boldsymbol{\beta})=n^{-1/2}\sum_{i=1}^{n}\int_{0}^{\infty}[\mathbf{q}(t)+\mathbf{b}\mathbf{\Omega}^{-1}\{\mathbf{W}_i(t)-\bar{\mathbf{w}}(t)\}]\,\mathrm{d}M_i(t)+o_p(1)$$

which in combination with (A1) yields $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})=n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\xi}_i+o_p(1)$, where

$$\boldsymbol{\xi}_i=\sum_{k=1}^{K}\frac{\delta_{ki}^{*}}{G(T_{ki}^{*}|\bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}+\int_{0}^{\infty}[\mathbf{q}(t)+\mathbf{b}\mathbf{\Omega}^{-1}\{\mathbf{W}_i(t)-\bar{\mathbf{w}}(t)\}]\,\mathrm{d}M_i(t)$$

Since $(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)$ are independent zero-mean random vectors, the multivariate central limit theorem entails that $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mathbf{V}=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\boldsymbol{\xi}_i^{\otimes 2}$.

For the Kaplan–Meier estimator

$$\frac{n^{1/2}\{G(t)-\hat{G}(t)\}}{G(t)}=n^{-1/2}\sum_{i=1}^{n}\int_{0}^{t}\frac{\mathrm{d}M_i(x)}{\pi(x)}+o_p(1)$$

where $\pi(t)=\Pr(X_i\geqslant t)$ $(i=1,\ldots,n)$, $M_i(t)=\bar{\delta}_i I(X_i\leqslant t)-\int_0^t I(X_i\geqslant x)\lambda(x)\,\mathrm{d}x$, and $\lambda(t)=-\mathrm{d}\log G(t)/\mathrm{d}t$. The aforementioned asymptotic normality for $n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$ still holds, but with

$$\boldsymbol{\xi}_i=\sum_{k=1}^{K}\frac{\delta^*_{ki}}{G(T^*_{ki}|\bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}+\int_0^{\infty}\mathbf{q}(t)\,\mathrm{d}M_i(t)$$

where

$$\mathbf{q}(t)=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{I(T^*_{ki}>t)\delta^*_{ki}h(\mathbf{Z}_{ki};\boldsymbol{\beta})\{y_{ki}-g(\boldsymbol{\beta}'\mathbf{Z}_{ki})\}\mathbf{Z}_{ki}}{G(T^*_{ki}|\bar{\mathbf{F}}_i)\pi(t)}$$

The Taylor series expansion of $\mathbf{U}(\hat{\boldsymbol{\beta}})$ at $\mathbf{U}(\boldsymbol{\beta})$ yields

$$n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\tilde{\mathbf{A}}^{-1}(\boldsymbol{\beta}^*)n^{-1/2}\mathbf{U}(\boldsymbol{\beta})$$

where $\tilde{\mathbf{A}}(\boldsymbol{\beta})=-n^{-1}\partial\mathbf{U}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$, and $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. The law of large numbers, together with the consistency of $\hat{\boldsymbol{\beta}}$ and $\hat{G}$, implies that $\tilde{\mathbf{A}}(\boldsymbol{\beta}^*)$ converges in probability to a deterministic matrix, say $\mathbf{A}$. Therefore, $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mathbf{A}^{-1}\mathbf{V}\mathbf{A}^{-1}$. The replacements of the unknown quantities in $\boldsymbol{\xi}_i$ with their sample estimators yield $\hat{\boldsymbol{\xi}}_i$ given in Section 2.2. The consistency of the covariance matrix estimator for $\hat{\boldsymbol{\beta}}$ follows from the law of large numbers, together with the consistency of $\hat{G}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$.

REFERENCES

1. Potosky AL, Riley GF, Lubitz JD, Mentnech RM, Kessler LG. Potential for cancer related health services research using a linked Medicare-tumor registry data base. *Medical Care* 1993; **31**:732–747.
2. Lin DY, Etzioni R, Feuer EJ, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997; **53**:419–434.
3. Lin DY. Linear regression of censored medical costs. *Biostatistics* 2000; **1**:35–47.
4. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
5. Little RJA. Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**:1112–1121.
6. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
7. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd edn. Chapman & Hall: New York, 1989.
8. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**:663–685.
9. Koul H, Susarla V, van Ryzin J. Regression analysis with randomly right-censored data. *Annals of Statistics* 1981; **9**:1276–1288.
10. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*: *Methodological Issues*, Jewell NP, Dietz K, Farewell VT (eds). Birkhäuser: Boston, MA, 1992; 297–331.
11. Lin DY, Ying Z. A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* 1993; **80**:573–581.
12. Zhao H, Tsiatis AA. A consistent estimator for the distribution of quality-adjusted survival time. *Biometrika* 1997; **84**:339–348.

13. Breslow NE. Contribution to the discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* 1972; **34**:216–217.
14. Etzioni RD, Urban N, Baker M. Estimating the costs attributable to a disease with application to ovarian cancer. *Journal of Clinical Epidemiology* 1996; **49**:95–103.
15. Cook RJ, Lawless JF. Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* 1997; **16**:911–924.
16. Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika* 2000; **87**:329–343.
17. Lin DY, Wei LJ, Ying, Z. Model-checking techniques based on cumulative residuals. *Biometrics* 2002; **58**:1–12.
18. Lin DY. Proportional means regression for censored medical costs. *Biometrics* 2000; **56**:775–778.
19. Therneau TM, Grambsch PM. *Modelling Survival Data*: *Extending the Cox Model.* Springer: New York, 2000.
20. Willan A, O'Brien, BJ. Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem. *Health Economics* 1996; **5**:297–305.
21. Willan A. Analysis, sample size and power for estimating incremental net health benefit from clinical trial data. *Controlled Clinical Trials* 2001; **22**:228–237.
22. Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps; a nonparametric approach to confidence interval estimation. *Health Economics* 1997; **6**:327–340.
23. O'Hagan A, Stevens JW, Montmartin J. Baysian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine* 2001; **20**:733–753.
24. Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994; **81**:73–81.