The performance of unadjusted and adjusted medical cost estimators – an application in colorectal cancer

Yi-Ting Hwang*, Wan-Lin Yeh
Department of Statistics
National Taipei University, Taipei, Taiwan
Hsun-Chih Kuo
Department of Statistics
National Chengchi University, Taipei, Taiwan

May 14, 2011

Abstract

Owing to aging, how to sufficiently use the limited resources becomes controversial. Knowing how to estimate medical cost accurately and efficiently becomes an important issue. However, it is common to have dropouts when analyzing medical cost. The naïve estimators ignoring the unobservable data may be biased. Lin et al. (1997) suggested partitioning the study duration and estimating the cost of each interval and then constructing the estimate by summing up the cost from each interval. Furthermore, to take into account of the unobservable data, Lin et al. (1997) and Band and Tsiatis (2000) proposed weighted estimators that used the survival probability and uncensored probability as the weight, respectively.

Furthermore, the medical cost may be related to many covariates. Baser et al. (2006) suggested using the general linear model for the longitudinal data to model the partitioned cost, where a random intercept is included. This paper extends the model to a more general parametric model. Furthermore, similar to the weighting concept in Lin et al. (1997), we suggest using the survival probability as the weight adjustment which is estimated by the Cox proportional hazards model. Simulations are used to evaluate the performance of the unadjusted and adjusted cost estimators under various scenarios. Finally, the proposed model is implemented on the data extracted from Health Insurance database for patients with the colorectal cancer.

*Corresponding author. Email: hwangyt@gm.ntpu.edu.tw

KEY WORDS: Medical cost, Inverse probability weighted method, Mixed model, Proportional hazards model

1 Introduction

Studying the cost of health care has been a heightened interest recently. One important issue of this is to find a more cost-effective assessment for treating a disease or develop a more standardized and efficient medical intervention. The data for such analysis are available from clinical trials, disease registries, health insurance records, etc. A common characteristic of such data is that some patients may not be followed completely to the end of the event of interest. Assuming the lifetime cost is of interest, then the time to the end of death, i.e. the lifetime, may be censored and the lifetime cost is also censored. As a result, the mean cost is often estimated using the average total cost from all available data, even though some of these data may be censored and some of these data may be completely observed. This may lead to erroneous inference.

A fundamental difference between censored survival time and censored lifetime cost exists as pointed out in Lin et al. (1997). Patients may accumulate the lifetime cost differently. A patient who accrues costs at higher rates tends to generate larger total costs at the end of the event, which means the total cost at the end of the event may be positively correlated with the total cost at the end of the censored event. This means that the censored total cost cannot be estimated based on the standard survival methodologies which often assume the independence between the survival time and the censoring time.

The early investigation of statistical models for analyzing the total cost only focuses primarily on the estimation of total cost without adjusting the covariates. To take into account of censoring cost, Lin et al. (1997) suggested partitioning the entire observing period into several fixed intervals and then estimating the average cost by the survival probability in each time interval multiplied by the sample mean of the total costs from that interval, whereas Bang and Tsiatis (2000) proposed using the inverse probability weighted methods to adjust the censored cost. Recently, the investigation for analyzing the total cost uses the parametric models such as the regression (Lin et al., 2000) and the mixed model (Baser et al., 2006) to incorporate the important covarates.

In this paper, we discuss various unadjusted cost estimates in Section 2. Section 3 extends the model proposed by Baser et al. (2006) and incorporates the survival probability computed from the proportional hazards model as the inverse probability weight to find the adjusted cost estimator. Section 4 evaluates the performance of various unadjusted cost estimators and adjusted cost estimators using Monte Carlo simulation. Section 5 illustrates the usage

of these estimators using a real data obtained from National Health Insurance Research Database. Discussions are given in Section 6.

2 Unadjusted total cost estimation

Let X_i and C_i denote the survival time and censoring time of the ith patient and be assumed to be independent. Also, let $T_i = \min(X_i, C_i)$ and $\delta_i = I[X_i \leq C_i]$, where $I[A]$ denotes the indicator of the event A. Assume the research duration be $[0, \tau)$, where τ is a pre-specified constant. Let the duration be partitioned into K intervals, where the kth interval is denoted as $[a_k, a_{k+1})$ and $a_1 = 0$, $a_{K+1} = \tau$. Let Y_{ik} denote the true accumulative cost for the kth interval for the ith subject and $\tilde{Y}_{ik}$ denote the observed accumulative cost for the kth interval for the ith subject Then the true total cost for the ith patient becomes

$$Y_i = \sum_{k=1}^{K} Y_{ik}.$$

Owing to censoring, some $\tilde{Y}_{ik}$ may be incomplete observed. Lin et al. (1997) used the following ways to record the cost for the given interval:

1. For $\delta_i = 1$ or $T_i = \tau$, then $\tilde{Y}_{ik} = Y_{ik}$ and the accumulative cost for the i subject is $Y_i = \sum_{k=1}^{K_i} Y_{ik}$, where $T_i \in [a_{K_i}, T_i)$.

2. When $T_i < \tau$ and T_i falls in $[a_{K_i}, a_{K_i+1})$, then

   (a) the accumulative cost for K_i intervals equals $\sum_{k=1}^{K_i-1} Y_{ik}$.

   (b) the cost for the K_i interval equals the cost $\tilde{Y}_{iK_i}$ accumulated in $[a_{K_i}, T_i]$.

   (c) and The cost beyond the K_i interval equals 0, when $\delta_i = 1$ and equals missing when $\delta_i = 0$.

Based on the preceding recording for the cost, a subject who is censored owing to the length of research would have a complete cost history. Thus, we need another indictor for those subject, which is denoted as $\eta_i = I[T_i = \tau]$. A hypothetical data listed in table 1 is used to demonstrate the preceding assumption.

Using the conditional expectation properties, Lin et al. (1997) proposed some cost estimators that are based on either uncensored samples or complete samples. Bang and Tsiatis (2000) suggested using the inverse probability weighted method to adjust the censored cost to derive a cost estimator. Section 2.1 reviews estimators proposed by Lin et al. (1997), while Section 4 describes estimates proposed by Bang and Tsiatis (2000). Their performances will be illustrated using Monte Carlo simulation in Section 4.

Table 1: Cost construction of $(\tilde{Y}_{ik},\ \eta_i)$ for hypothetical cost data and survival time

Individuals	Partitions 1 [0,1)	2 [1,2)	3 [2,3)	4 [3,4)	5 [4,5)	6 [5,6)	7 [6,7)	Y_i	T_i	δ_i	η_i
1	7	.	.	.	.	.	.	7	0.5	0	0
2	13	8	9	14	5	0	0	49	4.3	1	1
3	12	7	10	8	7	8	9	61	7	0	1
4	13	9	11	13	0	0	0	46	3.8	1	1
5	10	8	7	8	6	9	0	48	6	0	0
6	25	30	19	0	0	0	0	74	2.2	1	1

2.1 Unadjusted estimates by the conditional expectation

Let μ and μ_k denote the population average total cost and the population average total cost for the kth interval. Then, based on the conditional expectation property, we have

$$\mu = \sum_{k=1}^{K} \mu_k = \sum_{k=1}^{K} \mathrm{E}[Y_{ik}|T \geq a_k]P[T \geq a_k] + \sum_{k=1}^{K} \mathrm{E}[Y_{ik}|T < a_k]P[T < a_k] = \sum_{k=1}^{K} S_k E_k, \quad (1)$$

where $S_k = P[T \geq a_k]$ and $E_k = \mathrm{E}[Y_{ik}|T \geq a_k]$. Owing the censoring, E_k can be estimated based on the full sample or the uncensored sample.

When the full sample is used, the estimate of E_k uses all available data in the kth interval. Let $\eta_{ik}^{\mathrm{FH}} = I[T_i \geq a_k]$. Then, an estimate of E_k can be obtained as

$$\hat{E}_k^{\mathrm{FH}} = \frac{\sum_{i=1}^{n} \eta_{ik}^{\mathrm{FH}} \tilde{Y}_{ik}}{\sum_{i=1}^{n} \eta_{ik}^{\mathrm{FH}}}, k = 1, \ldots, K.$$

When the uncensored sample is used, only the complete data in the kth interval is used to compute the estimate of E_k. Let $\eta_{ik}^{\mathrm{UH}} = I[T_i \geq a_k, C_i \geq \min(X_i, a_{k+1})]$, where $\eta_{iK}^{\mathrm{UH}} = 1$ if $T_i = \tau$. The estimate becomes

$$\hat{E}_k^{\mathrm{UH}} = \frac{\sum_{i=1}^{n} \eta_{ik}^{\mathrm{UH}} \tilde{Y}_{ik}}{\sum_{i=1}^{n} \eta_{ik}^{\mathrm{UH}}}, k = 1, \ldots, K.$$

Table 2 illustrates how to compute the estimates of $\hat{E}_k^{\mathrm{FH}}$ and $\hat{E}_k^{\mathrm{UH}}$. Obviously, $\hat{E}_k^{\mathrm{FH}}$ would be smaller than $\hat{E}_k^{\mathrm{UH}}$ when there are censored observations in the kth interval, while when no censored observations are found, we have $\hat{E}_k^{\mathrm{FH}}$=$\hat{E}_k^{\mathrm{UH}}$

By again the conditional expectation property, (1) can be re-expressed as

$$\mu = \sum_{k=1}^{K+1} \mathrm{E}[Y_i|a_k \leq T < a_{k+1}]P[a_k \leq T < a_{k+1}] = \sum_{k=1}^{K+1} (S_k - S_{k+1})A_k, \quad (2)$$

Table 2: Estimates of $\tilde{Y}_{ik}$, $\hat{E}_k^{\mathrm{FH}}$ and $\hat{E}_k^{\mathrm{UH}}$ using hypothetical data in Table 1

Individuals	Partitions 1 [0,1)	2 [1,2)	3 [2,3)	4 [3,4)	5 [4,5)	6 [5,6)	7 [6,7)	Y_i	T_i	δ_i	η_i
1	7	.	.	.	.	.	.	7	0.5	0	0
2	13	8	9	14	5	0	0	49	4.3	1	1
3	12	7	10	8	7	8	9	61	7	0	1
4	13	9	11	13	0	0	0	46	3.8	1	1
5	10	8	7	8	6	9	0	48	6	0	0
6	25	30	19	0	0	0	0	74	2.2	1	1
$\sum_{i=1}^{6} \eta_{ik}^{\mathrm{FH}}$	6	5	5	4	3	2	2				
$\hat{E}_k^{\mathrm{FH}}$	15	12.4	11.2	10.75	6	17	4.5				
$\sum_{i=1}^{n} \eta_{ik}^{\mathrm{UH}}$	5	5	5	4	3	2	1				
$\hat{E}_k^{\mathrm{UH}}$	14.6	12.4	11.2	10.75	6	17	9				

where $A_k = \mathrm{E}[Y_i | a_k \leq T < a_{k+1}]$ and $a_{K+2} = \infty$. Let $\eta_{ik}^{\mathrm{T}} = I[a_k \leq T_i < a_{k+1}]$, where $\eta_{iK}^{\mathrm{T}} = 1$ if $T_i = \tau$. Using all possible sample in the interval, an estimate of A_k can be

$$\hat{A}_k = \frac{\sum_{i=1}^{n} \eta_{ik}^{\mathrm{T}} \tilde{Y}_i}{\sum_{i=1}^{n} \eta_{ik}^{\mathrm{T}}}, k = 1, \ldots, K.$$

Table 3 illustrates how to compute the estimates of $\hat{A}_k^{\mathrm{T}}$. Furthermore, the survival S_k in (1) and (2) can be estimated by the Kaplan-Meier estimator (Kaplan and Meier, 1958) which is denoted as $\hat{S}_k$. Based on replacing $\hat{E}_k^{\mathrm{FH}}$, $\hat{E}_k^{\mathrm{UH}}$ and $\hat{A}_k$ and S_k in (1) and (2), we can obtain three estimates of cost and be denoted as $\hat{\mu}^{\mathrm{FH}}$, $\hat{\mu}^{\mathrm{UH}}$, $\hat{\mu}^{\mathrm{T}}$, respectively. The asymptotic properties of $\hat{\mu}^{\mathrm{FH}}$, $\hat{\mu}^{\mathrm{UH}}$, $\hat{\mu}^{\mathrm{T}}$ are discussed in Lin et al. (1997).

2.2 Unadjusted estimators by IPWM

Cost estimators that are based on the conditional expectation properties do not consider the censored cost. To take into account of censored cost, Bang and Tsiatis (2000) proposed using inverse probability weighted method (IPWM) to adjust for the censored cost. The detail of IPWM is referred to Carpenter and Kenward (2005). Let $p(t) = P[C_i \geq t]$ denote the uncensored probability at time t, which can be estimated again by Kaplan-Meier estimator. Let this estimator be denoted as $\hat{p}(t)$. To avoid computational complexity, we use the conventional means to define $\hat{p}(t)$ at the last observation, that is, the last observation is treated as a censored case regardless the actual status is.

Table 3: Estimates of $\tilde{Y}_{ik}$, $\hat{E}_k^{\mathrm{T}}$ using hypothetical data in Table 1

Individuals	Partitions 1 [0,1)	2 [1,2)	3 [2,3)	4 [3,4)	5 [4,5)	6 [5,6)	7 [6,7)	Y_i	T_i	δ_i	η_i
1	7	.	.	.	.	.	.	7	0.5	0	0
2	13	8	9	14	5	0	0	49	4.3	1	1
3	12	7	10	8	7	8	9	61	7	0	1
4	13	9	11	13	0	0	0	46	3.8	1	1
5	10	8	7	8	6	9	0	48	6	0	0
6	25	30	19	0	0	0	0	74	2.2	1	1
$\sum_{i=1}^{n} \eta_{ik}^{\mathrm{T}}$	0	0	1	1	1	0	1				
$\hat{A}_k$	0	0	74	46	49	0	61				

Since

$$E\left[\frac{\delta_i}{p(T_i)}\right] = E\left\{E\left[\frac{\delta_i}{p(T_i)}\right]|T_i\right\} = 1, \tag{3}$$

a weighted estimator using only uncensored observation can be given by

$$\hat{\mu}^{\mathrm{BT}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\eta_i \tilde{Y}_i}{\hat{p}(T_i)} \tag{4}$$

The estimator defined in (3) uses only the total cost and all the cost history are not included. Adapted the partition idea by Lin et al. (1997), Bang and Tsiatis (2000) then proposed a partitioned weighted estimator defined as

$$\hat{\mu}^{\mathrm{BTP}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\eta_{ik}^{\mathrm{BTP}}\tilde{Y}_{ik}}{\hat{p}(T_{ik})},$$

where $\eta_{ik}^{\mathrm{BTP}} = I[C_i \geq T_{ik}]$, $T_{ik} = \min(T_i, a_{k+1})$ and $\eta_{iK}^{\mathrm{BTP}} = 1$ if $T_i = \tau$. The asymptotic properties of $\hat{\mu}^{\mathrm{BT}}$ and $\hat{\mu}^{\mathrm{BTP}}$ are discussed in Bang and Tsiatis (2000).

3 Adjusted cost estimations

The medical cost often varies dramatically with the disease severity, the type of treatment and the level of ward, etc. Thus, the model-based estimator should be considered. However, owing to censoring, some costs are not observable. Furthermore, when the cost history is

considered, there may exist correlation among cost history. Some adjustments in the ordinary regression model are needed.

To simplify the notation, let

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iK} \end{pmatrix},$$

and let the covariates for the i subject be denoted as

$$\boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{Z}_{i1} \\ \boldsymbol{Z}_{i2} \\ \vdots \\ \boldsymbol{Z}_{iK} \end{pmatrix} = \begin{pmatrix} 1 & Z_{i11} & Z_{i12} & \cdots & Z_{i1,p-1} \\ 1 & Z_{i21} & Z_{i22} & \cdots & Z_{i1,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{iK1} & Z_{iK2} & \cdots & Z_{iK,p-1} \end{pmatrix},$$

where $\boldsymbol{Z}_{ij}$ is the covariate observed at the jth interval. Define the linear model for the repeated measurements as

$$Y_{ik} = \beta_0 + \beta_1 Z_{ik,1} + \ldots + \beta_{p-1} Z_{ik,p-1} + \epsilon_{ik}, i = 1, \ldots, n, j = 1, \ldots, K \tag{5}$$

or $\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, i = 1, \ldots, n$, where $\boldsymbol{\beta}^{'} = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ and $\boldsymbol{\epsilon}_i^{'} = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iK})$ is the measurement error. The distribution of $\boldsymbol{\epsilon}_i$ are assumed to be multivariate normal with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_i$.

The maximum likelihood estimation can be used to estimate $(\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$ and let the estimates be denoted as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{Z}_i^{'} \hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i^{'} \hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Y}_i \right) \tag{6}$$

and $\hat{\boldsymbol{\Sigma}}_i$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is the mutlivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}, \tag{7}$$

where $\boldsymbol{A} = \sum_{i=1}^{n} \boldsymbol{Z}_i^{'}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i$, $\boldsymbol{B} = \sum_{i=1}^{n} \boldsymbol{Z}_i^{'}\boldsymbol{\Sigma}_i^{-1}\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i^{'}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i$, $\hat{\boldsymbol{u}}_i = \boldsymbol{Y}_i - \boldsymbol{Z}_i\boldsymbol{\beta}$.

When $\boldsymbol{\Sigma}_i$ is completely unspecified and the number of partitions increases, the number of parameters in $\boldsymbol{\Sigma}_i$ increases dramatically. Since there may exist certain association characteristics for the repeated measures, $\boldsymbol{\Sigma}_i$ is often characterized into a specified parametric structures such as the compound symmetry structure, first-order autoregressive structure, Toeplitz structure, etc (see Fitzmaurice, Laird and Ware, 2004). An appropriate structure can be determined by the likelihood ratio statistics or Akaike information criterion.

Model (5) is no longer appropriate when data are unbalanced. The mixed model that includes the fixed effect and random effect, can be then used. Define the mixed model as

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \boldsymbol{D}_i\boldsymbol{u}_i + \boldsymbol{e}_i, i = 1, \ldots, n, \tag{8}$$

where $\boldsymbol{D}_i$ is the design matrix for random effects, which is a subset of $\boldsymbol{Z}_i$, and $\boldsymbol{u}_i$ is the vector for random effects. Under model (8), $\boldsymbol{u}_i$ is assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{G}$ and be independent of $\boldsymbol{e}_i$ and $\boldsymbol{e}_i$ a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{R}_i$. In practice, owing to limited data, $\boldsymbol{R}_i$ is assumed to be a diagonal matrix. Using transformation, we can obtain $\boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^{'} + \boldsymbol{R}_i$. The maximum likelihood estimation can be again used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_i$.

Baser et al. (2006) suggested using the random intercept model to estimate the effects of treatment on the total medical cost which is weighted by IPWM. The following extends the random intercept model to a more general settings. Two situations are included:

1. The censoring bias is not adjusted.
2. The censoring bias is adjusted using IPWM.

3.1 Scenario I

Owing to censoring, some cost data are unobservable. To estimate the effects of treatment on the total medical cost, we modify the design matrix $\boldsymbol{Z}_i$ as $\tilde{\boldsymbol{Z}}_i = \boldsymbol{S}_i\boldsymbol{Z}_i$, which is the design matrix for the observable cost, where $\boldsymbol{S}_i$ is a $K \times K$ diagonal matrix and the kth diagonal element equals 1 when the cost for the kth interval for the ith subject is observable and 0 otherwise. Based on only observable data, the model defined in (5) becomes

$$\tilde{\boldsymbol{Y}}_i = \tilde{\boldsymbol{Z}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{e}}_i, i = 1, \ldots, n, \tag{9}$$

where $\tilde{\boldsymbol{e}}_i$ has a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_i$. An important assumption is needed when using this model. That is, we need to assume that censoring is not associated with the total cost.

Using the maximum likelihood estimation, we can obtain the maximum likelihood estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_I = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{'}\boldsymbol{\Sigma}_i^{-1}\tilde{\boldsymbol{Z}}_i\right)^{-1}\left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{'}\boldsymbol{\Sigma}_i^{-1}\tilde{\boldsymbol{Y}}_i\right) \tag{10}$$

and $\hat{\boldsymbol{\Sigma}}_i^{-1}$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}_I$ is multivariate normal with mean $\boldsymbol{\beta}$ and the asymptotic variance covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I) = \hat{\boldsymbol{A}}_I^{-1}\hat{\boldsymbol{B}}_I\hat{\boldsymbol{A}}_I^{-1}$, where $\hat{\boldsymbol{A}}_I = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{'}\boldsymbol{\Sigma}_i^{-1}\tilde{\boldsymbol{Z}}_i\right)$, $\hat{\boldsymbol{B}}_I = \sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{'}\boldsymbol{\Sigma}_i^{-1}\hat{\tilde{\boldsymbol{e}}}_i\hat{\tilde{\boldsymbol{e}}}_i'\boldsymbol{\Sigma}_i^{-1}\tilde{\boldsymbol{Z}}_i$, $\hat{\tilde{\boldsymbol{e}}}_i = \tilde{\boldsymbol{Y}}_i - \tilde{\boldsymbol{Z}}_i\boldsymbol{\beta}_I$.

3.2 Scenario II

The estimator defined in (10) is derived based on only the observable data. To avoid such assumption, we use the IPWM to adjust the missing cost. Let $\boldsymbol{P}_i$ denote the uncensored

probability for the ith subject, where the kth diagonal element is $\sqrt{p_{ik}}$ and p_{ik} is the uncensored probability for the kth interval for the ith subject. Based on IPWM, the design matrix and cost are modified as $\tilde{\boldsymbol{Z}}_i^P = \boldsymbol{P}_i^{-1}\tilde{\boldsymbol{Z}}_i$, $\tilde{\boldsymbol{Y}}_i^P = \boldsymbol{P}_i^{-1}\tilde{\boldsymbol{Y}}_i$ and the corresponding model becomes

$$\tilde{\boldsymbol{Y}}_i^P = \tilde{\boldsymbol{Z}}_i^P\boldsymbol{\beta} + \tilde{\boldsymbol{e}}_i^P, i = 1, \ldots, n. \tag{11}$$

The estimator of $\boldsymbol{\beta}$ can be obtained using the maximum likelihood method as

$$\hat{\boldsymbol{\beta}}_{II} = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{P'} \boldsymbol{\Sigma}_i^{-1} \tilde{\boldsymbol{Z}}_i^P\right)^{-1} \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{P'} \boldsymbol{\Sigma}_i^{-1} \tilde{\boldsymbol{Y}}_i^P\right).$$

The corresponding asymptotic variance-covariance matrix is $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{II}) = \hat{\boldsymbol{A}}_{II}^{-1}\hat{\boldsymbol{B}}_{II}\hat{\boldsymbol{A}}_{II}^{-1}$, where $\hat{\boldsymbol{A}}_{II} = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{P'} \boldsymbol{\Sigma}_i^{-1} \tilde{\boldsymbol{Z}}_i^P\right)$, $\hat{\boldsymbol{B}}_{II} = \sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i^{P'} \boldsymbol{\Sigma}_i^{-1} \hat{\tilde{\boldsymbol{e}}}_i^P \hat{\tilde{\boldsymbol{e}}}_i^{P'} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Z}_i^P$, $\hat{\tilde{\boldsymbol{e}}}_i^P = \tilde{\boldsymbol{Y}}_i^P - \tilde{\boldsymbol{Z}}_i^P\hat{\boldsymbol{\beta}}_{II}$.

4 Simulation

The performance of the unadjusted cost estimators and adjusted cost estimators is evaluated using Monte Carlo simulations. Simulated data include two parts of information, one is the information for the survival information and the other one is the information about the cost.

The simulation scheme is a modification of Lin et al. (2000). The survival times is generated from either the exponential distribution with mean $c = 6$ years or the uniform distribution (0,10) years, while the censoring time is generated from the uniform distribution on $(0, c)$ years, where c is determined according to the probability of censoring. We consider three censoring situations, 30%, 40% and 50%. Under the exponential survival, c equals 20, 12.5 and 10, while under the uniform survival, c equals 17, 12.5 and 10. The simulation setting for cost is described separately for the unadjusted and adjusted estimators as follows.

The performance of each estimator is evaluated based on the bias, the standard error of estimator (SSE), the mean of standard error of estimator (SEE) and the 95% coverage probability (CP), which are computed from 10,000 simulated samples.

4.1 Performance of unadjusted cost estimations

The total cost includes three types of cost, the baseline cost, the diagnostic cost and the cost in the final year of life (cost of mortality). We assume the total duration τ equals 10 years and the costs are observed each year, that is, $K = 10$. Within each year, there is a baseline cost of uniform (100,300). Additionally, there is a diagnostic cost of uniform (500,1500) at $t = 0$ and a cost of final year of life of uniform (1000,3000). Based on the preceding setting, for the uniform survival, the total cost μ equals \$39, 000, whereas for the exponential survival, the total cost μ equals \$34, 380. Finally, the sample size is assumed to be 100, 250 and 500.

Tables 4 and 5 summary the results of simulation studies on the unadjusted estimates discussed in Section 2. The bias tends to be larger when using the conditional properties to construct the estimators. In particular, owing to the construction, $\hat{\mu}^{\mathrm{FH}}$ using only the full sample the tends to underestimate the total cost, while $\hat{\mu}^{\mathrm{UH}}$ using only uncensored sample tends to overestimate the total cost. On the other hand, the results of simulation also demonstrates that the performance of $\hat{\mu}^{\mathrm{T}}$ and $\hat{\mu}^{\mathrm{BT}}$ is similar. $\hat{\mu}^{\mathrm{BTP}}$ has the smallest bias among most of simulated situations. The environmental settings can influence the bias. As n increases, the bias becomes smaller, except for $\hat{\mu}^{\mathrm{UH}}$. The impact of n is especially strong on $\hat{\mu}^{\mathrm{T}}$, $\hat{\mu}^{\mathrm{BT}}$ and $\hat{\mu}^{\mathrm{BTP}}$. On the other hand, the percent of censoring has negative impact on the bias. The more the censoring observations, the larger the bias. In particular, as compared to other estimates, $\hat{\mu}^{\mathrm{BTP}}$ has the least impact. Since $\hat{\mu}^{\mathrm{T}}$ and $\hat{\mu}^{\mathrm{BT}}$ use only one weight for each subject, these estimates have large bias as the percent of censoring increases. Furthermore, the bias of $\hat{\mu}^{\mathrm{BTP}}$ is robust with respect to the distribution of survival times, while the bias of the other estimates is not. Since $\hat{\mu}^{\mathrm{UH}}$ uses the uncensored sample and $\hat{\mu}^{\mathrm{T}}$ and $\hat{\mu}^{\mathrm{BT}}$ use only one weight for each subject, when the distribution of survival is right-skewed, the bias becomes extremely large. The influence of distribution is even more evident when the percent of censoring increases.

The performance in terms of SSE and SEE is similar for five estimates. As expected, SSE and SEE decrease as the sample size increases, while increasing the censoring probability enlarges SSE and SEE. The performance in SSE and SEE is very robust with respect to the distribution. However, when the censoring probability is more than 40%, SSE and SEE increase dramatically.

The performance in terms of CP for $\hat{\mu}^{\mathrm{T}}$, $\hat{\mu}^{\mathrm{BT}}$ and $\hat{\mu}^{\mathrm{BTP}}$ is good when the censoring probability is less than 50%, while CP can become lower than 70% when the censoring probability equals 50%. In particular, when the survival distribution is right-skewed, CP for $\hat{\mu}^{\mathrm{T}}$ and $\hat{\mu}^{\mathrm{BT}}$ can be lower than 30%. Furthermore, increasing the sample size increases CP, except when the probability of censoring is more than 40%. Surprisingly, CP of $\hat{\mu}^{\mathrm{FH}}$ and $\hat{\mu}^{\mathrm{UH}}$ is influenced differently by the sample size and distribution. To be more specifically, increasing the sample size deduces CP, while CP is larger for $\hat{\mu}^{\mathrm{FH}}$ and $\hat{\mu}^{\mathrm{UH}}$ when the survival distribution is skewed.

4.2 Performance of adjusted cost estimations

To evaluate the adjusted cost estimators, besides the earlier setting, a random intercept is added to characterize the subject heterogeneity. We assume the random intercept has a uniform distribution on $(0, c)$, where c is determined by the intraclass correlation coefficient (ICC). The ICC is defined as $\mathrm{ICC} = \sigma_u^2/(\sigma_e^2 + \sigma_u^2)$, where σ_u^2 is the variance of the random

Table 4: Performance of unadjusted cost estimator under the uniform survival

$P[C \leq X]$	c	n		$\hat{\mu}^{\mathrm{FH}}$	$\hat{\mu}^{\mathrm{UH}}$	$\hat{\mu}^{\mathrm{T}}$	$\hat{\mu}^{\mathrm{BT}}$	$\hat{\mu}^{\mathrm{BTP}}$
30%	17	100	Bias	-1179	505	-61	-40	-16
			SSE	1143	1181	1193	1192	1153
			SEE	1102	1134	1144	1018	1044
			CP*	80.3%	91.3%	93.8%	90.3%	92.1%
		250	Bias	-1173	510	-27	-4	0
			SSE	723	741	739	739	726
			SEE	709	730	733	647	654
			CP	61.7%	88.9%	94.4%	91.1%	91.9%
		500	Bias	-1184	501	-34	-10	-8
			SSE	504	521	518	518	508
			SEE	504	520	520	458	461
			CP	35.3%	83.9%	95.0%	91.8%	92.3%
40%	12.5	100	Bias	-1990	800	-286	-255	-121
			SSE	1284	1359	1553	1554	1313
			SEE	1201	1250	1256	1043	1061
			CP	61.5%	88.4%	90.3%	83.8%	88.5%
		250	Bias	-1980	852	-108	-69	-34
			SSE	795	826	857	858	798
			SEE	774	807	805	653	658
			CP	28.2%	81.4%	93.4%	87.1%	89.3%
		500	Bias	-1971	865	-56	-16	-7
			SSE	552	576	573	575	549
			SEE	550	576	569	459	462
			CP	5.2%	68.2%	94.4%	88.5%	90.1%
50%	10	100	Bias	-3866	497	-2831	-2793	-1224
			SSE	1758	2042	3598	3600	1957
			SEE	1425	1508	1781	1386	1058
			CP	28.2%	83.7%	60.7%	51.3%	62.8%
		250	Bias	-3716	1118	-1802	-1750	-765
			SSE	1099	1325	2289	2298	1267
			SEE	961	1015	1209	920	650
			CP	5.2%	73.4%	64.9%	53.7%	62.0%
		500	Bias	-3687	1435	-1279	-1221	-531
			SSE	750	903	1644	1652	918
			SEE	711	736	895	670	457
			CP	0.1%	47.7%	66.7%	54.5%	60.8%

* CP is the coverage probability of the 95% confidence interval

Table 5: Performance of unadjusted cost estimator under the exponential survival

$P[C \leq X]$	c	n		$\hat{\mu}^{\mathrm{FH}}$	$\hat{\mu}^{\mathrm{UH}}$	$\hat{\mu}^{\mathrm{T}}$	$\hat{\mu}^{\mathrm{BT}}$	$\hat{\mu}^{\mathrm{BTP}}$
30%	20	100	Bias	-762	358	-37	-18	-17
			SSE	1114	1175	1156	1156	1141
			SEE	1098	1154	1107	1124	892
			CP	88.1%	93.5%	93.5%	93.9%	87.0%
		250	Bias	-753	372	-27	-6	-5
			SSE	697	736	723	723	716
			SEE	697	733	717	712	597
			CP	80.5%	92.3%	94.5%	94.4%	89.6%
		500	Bias	-745	379	-18	3	2
			SSE	495	521	514	514	507
			SEE	493	519	511	504	429
			CP	66.6%	89.1%	94.7%	94.4%	90.1%
40%	14.1	100	Bias	-1239	634	-50	-23	-24
			SSE	1198	1339	1297	1296	1261
			SEE	1184	1306	1197	1296	955
			CP	79.4%	92.5%	92.2%	93.8%	85.9%
		250	Bias	-1227	640	-43	-13	-13
			SSE	760	846	821	821	800
			SEE	751	832	791	820	608
			CP	61.2%	88.6%	93.9%	94.9%	86.3%
		500	Bias	-1220	649	-41	-10	-7
			SSE	534	594	579	578	561
			SEE	532	590	567	579	431
			CP	38.2%	81.4%	94.7%	95.1%	86.8%
50%	9.6	100	Bias	-2716	874	-6494	-6452	-764
			SSE	1598	2607	4283	4287	1989
			SEE	1363	1598	2522	1776	939
			CP	45.2%	75.8%	30.3%	20.1%	59.1%
		250	Bias	-2676	1676	-6179	-6130	-587
			SSE	996	2461	3099	3107	1429
			SEE	919	1115	1975	1309	605
			CP	21.9%	64.9%	18.9%	11.5%	53.6%
		500	Bias	-2658	2686	-6008	-5956	-505
			SSE	694	2538	2405	2409	1119
			SEE	668	808	1617	1042	432
			CP	6.2%	45.3%	11.9%	7.9%	49.3%

* CP is the coverage probability of the 95% confidence interval

intercept and σ_e^2 is the variance of the measurement error. Three different ICC values are considered as 0.2, 0.5 and 0.7. The total cost is also classified into three types, the baseline cost, the diagnostic cost and mortality cost. All of these costs are generated from uniform (0,1). Based on the mean response curve, there is a sudden drop at the second year. Thus, the piecewise linear model defined as

$$Y_{ik} = \beta_0 + \beta_1 Z_i + \beta_2 t_{ik} + \beta_3 (t_{ik} - 2)_+ + u_i + e_{ik},$$

where Z_i is a dummy variable for the treatment. There is no treatment effect in the simulated data and thus the true value for β_1 is 1.

Table 6 and 7 list the result for simulations. Adjusted estimates are estimated from four models:

Model I $\mathbf{\Sigma}_i$ is a diagonal matrix and no adjustment for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}_I$.

Model II $\mathbf{\Sigma}_i$ is a diagonal matrix and IPWM adjustment is applied for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}_{II}$.

Model III Dependence is considered and no adjustment for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}_{III}$.

Model IV Dependence is considered and IPWM adjustment is applied for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}_{IV}$.

The bias for these estimates should be similar. The performance of these estimates should be compared in terms of SSE, SEE and CP. As expected, SSE and SEE for $\hat{\beta}_{III}$ and $\hat{\beta}_{IV}$ are smaller than $\hat{\beta}_I$ and $\hat{\beta}_{II}$. The difference in SSE and SEE between ($\hat{\beta}_{III}$, $\hat{\beta}_{IV}$) and ($\hat{\beta}_I$, $\hat{\beta}_{II}$) is actually governed by ICC. The larger ICC, the more dependence between data. As a result, as ICC increases, the difference in SSE and SEE between ($\hat{\beta}_{III}$, $\hat{\beta}_{IV}$) and ($\hat{\beta}_I$, $\hat{\beta}_{II}$) increases. Furthermore, increasing sample size increases the precisions of the estimates. In addition, as previous seen in unadjusted estimates, the proportion of censoring has very mild impact on SSE and SEE, whereas all the estimates are very robust with respect to the survival distribution. The performance in terms of CP is good for all four estimates.

5 Application to Colorectal cancer data

The medical cost for patients who had colorectal cancer extracted from the Health Insurance database in Taiwan was studied in this section. We considered 7646 patients who were diagnosed first with colorectal cancer in 2001. The data on mortality and yearly medical

costs were collected during the period of 2001 to 2006. three basic demographic variables, sex, age and area of residence, and two hospital characteristics, type of hospitals (private or public) and hospital level (regional hospital , teaching hospital and medical center) l were included. Age was categorized into 6 groups, under 35 years of age, 36-45, 46-55, 56-65, 65-75 and over 76. The area of residence included Taipei city, northern region, central region, southern region and Kaohsiung. Furthermore, three disease related variables, Charlson index, stage of cancer and treatment were recorded. We considered four treatment combinations, no treatment, surgery and chemotherapy, surgery only and chemotherapy and radiotherapy. It is important to understand how the treatment along with other covariates affects the medical cost and survival.

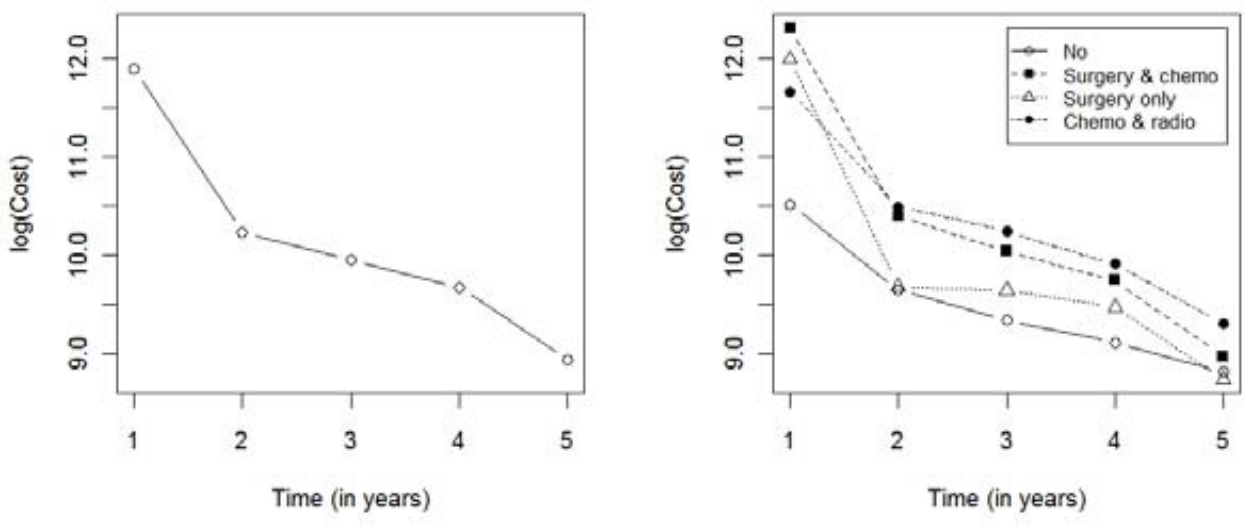
The survival time and medical costs are censored on the patients who were still alive at the end of 2006. The censoring was sorely caused by the limited study duration and it is reasonable to assume that censoring is independent of all other variables.

Table 8 provides the summary statistics of controlling variables and unadjusted cost estimates stratified by various controlling variables, where $\hat{\mu}^{\mathrm{BTP}}$ is used to compute the medical cost. There were 43under 35 years of age, 6.38% aged 36-45 years, 12.6% aged 46-55 years, 20.6% aged 56-65 years and 32.72% aged 66-75 years. 33.8% patients resided in Taipei city and 13% of patients lived in southern area. 67.5% of patients were treated in the private hospital and 51% of patients were treated in the regional hospital. Most of patients had stage I (61%), while only 22.6% of patients had stage III and IV. Furthermore, 56% of patients treated with surgery and chemotherapy and 14% of patients did not have any treatment.

The medical cost for male patients is higher (NT 50,000 dollars) than that for female patients. Patients who were under 55 years of age had higher cost (over NT 50,000 dollars) than those over 55 years of age. Patients lived in Taipei city would spend almost NT 50,000 dollars more than other areas, while patients lived in northern area had lowest medical cost. Patients who were treated in public hospitals and medical centers would have higher medical cost. The higher the disease stage, the higher the medical cost. Patients treated with surgery and chemotherapy had highest cost and those treated only chemotherapy and radiotherapy had the second higher medical cost.

Figure 2 displays the mean response curve for the medical cost. For the colorectal cancer, the medical cost for the first-year cost is much higher than that for the following years. To find an appropriate covariance pattern structure for the cost model, we use the 4th degrees of polynomial mean models with a random intercept controlling for sex, age group, residential areas, type of hospitals, hospital levels, cancer stage and type of treatments. Table 9 lists the result for the model of fit for various covariance-covariance pattern. The Ante-dependence structure (ANTE(1)), whose ijth element is $\sigma_i\sigma_j\prod_{k=i}^{j-1}\rho_k$, has the smallest AIC value for

both the unweighted and weighted model. The ANTE(1) is used to establish the final mean model and the model of fit is listed in Table 10. Based on the likelihood ratio test, the cubic model is selected.



(a) All sample (b) Weighted cost adjusted estimates

Figure 1: Stratified by treatment

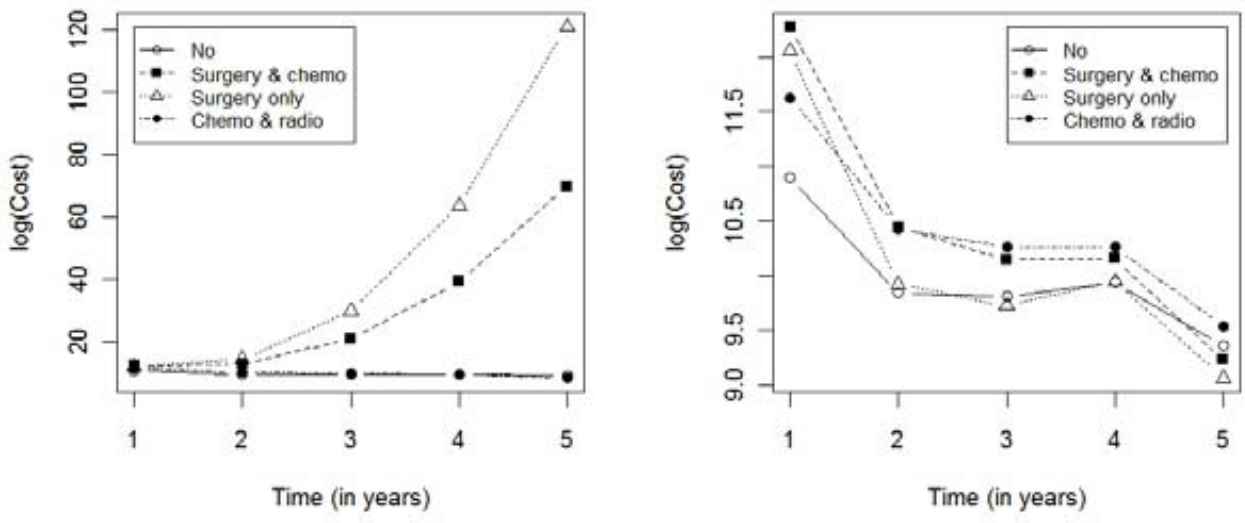
Figure 2: Mean response curve for medical cost

Tables 11 list the coefficient estimates for the unweighted model and weighted model. Most of parameter estimators obtained from these two models are similar, except for the age and time effect by treatment interaction.

As compared to patients with 76 years of age and older, the younger group (under 35 years old and 36-45 years of age) would spend significantly more medical expenditure. Since a cubic model is constructed, Figures 3 (a) and 3 (b) are drawn to demonstrate the difference in the parameter estimations of time effect by treatment interaction. Controlling for all other factors, the estimated medical cost in the unweighted model increases as the followup time increases, while the estimated medical cost in the weighted model increases. Since the trend displayed in Figure 3 (b) is similar to that showed in Figure 2 (b), the weighted model seems to have a better fit. Based on the weighted model, patients with no treatment have the least average medical cost in the first year, while patients with surgery and chemotherapy have highest average medical cost in the first year. For the second to fourth year, patients with surgery and chemotherapy and chemotherapy and radiotherapy spend similar, while average

costs for the other two groups are similar. At the fifth year, the highest medical cost becomes patients with chemotherapy and radiotherapy.

Controlling for other factors, males and patients who are 45 years of age or younger spend significantly more, Patients who reside in Taipei city have significantly higher averaged cost than those who reside in other areas. In particular, the largest difference appears between Taipei city and northern area. As expected, patients who are treated in medical center would spend more that patients who are treated in regional hospitals and teaching hospitals. Furthermore, patients who have higher stage and more comorbidity would have higher average medical cost.



(a) Unweighted cost adjusted estimates (b) Weighted cost adjusted estimates

Figure 3: Estimated medical cost based on the model in Table 11

6 Discussion and conclusion

The performance of the unadjusted estimate $\hat{\mu}^{\mathrm{BTP}}$ is good when the censoring probability is less than 50%. Although $\hat{\mu}^{\mathrm{FH}}$ and $\hat{\mu}^{\mathrm{UH}}$ uses cost histories to construct the estimates, the construction of these estimates did not actually correct the information of censoring cost. The estimate $\hat{E}_k^{FH}$ in $\hat{\mu}^{\mathrm{FH}}$ uses the full sample, which yields a smaller weight for each observation, while the estimate $\hat{E}_k^{UH}$ in $\hat{\mu}^{\mathrm{UH}}$ uses the uncensored sample, which yields a larger weight for each observation. As a result, the former estimate underestimates the cost, whereas the later overestimate the cost. Although the construction of $\hat{\mu}^{\mathrm{T}}$ and $\hat{\mu}^{\mathrm{BT}}$ is different, the underlying

concept is similar. That is, both estimates use the inverse of the subjects that are at risk as a weight. The difference in performance between the two estimates may be due that $\hat{\mu}^{\mathrm{T}}$ takes into account of cost history, while $\hat{\mu}^{\mathrm{BT}}$ uses the total cost.

When the probability of censoring is 50%, the accurate and precision decline dramatically, especially when the survival distribution is right-skewed. This may result from reducing the number of observable cost history data. Thus, other weight adjustments may be considered. For example, one may use the methodology proposed by Robins, Rotnitzky and Zhao (1995) and Hogan, Roy and Korkontzelou (2004) to handle the missing data in the longitudinal setting.

The linear model for repeated measures adjusted by the inverse probability method performs well under the simulation setting. In addition, a joint model proposed by Liu, Wolfe and Kalbfleish (2007) for the mixed model and proportional hazards model that include a shared random effect to incorporate the association between medical costs and survival may be considered.

Table 6: Performance of adjusted cost estimator under the uniform survival

			$P[C \leq X] = 0.3$				$P[C \leq X] = 0.4$			
n	ρ		$\hat{\beta}_I$	$\hat{\beta}_{II}$	$\hat{\beta}_{III}$	$\hat{\beta}_{IV}$	$\hat{\beta}_I$	$\hat{\beta}_{II}$	$\hat{\beta}_{III}$	$\hat{\beta}_{IV}$
100	0.2	Bias	0.0007	0.0007	0.0006	0.0006	0.0003	0.0004	0.0005	0.0006
		SSE	0.0463	0.0474	0.0465	0.0460	0.0489	0.0520	0.0496	0.0499
		SEE	0.0456	0.0463	0.0466	0.0461	0.0474	0.0491	0.0487	0.0485
		CP*	94.2%	94.0%	94.7%	94.7%	94.1%	93.4%	94.7%	94.2%
	0.5	Bias	-0.0017	-0.0017	-0.0009	-0.0010	0.0014	0.0008	0.0019	0.0017
		SSE	0.1025	0.1066	0.0952	0.0941	0.1068	0.1148	0.1001	0.0989
		SEE	0.1009	0.1041	0.0949	0.0938	0.1034	0.1094	0.0974	0.0963
		CP	94.2%	94.0%	94.4%	94.3%	93.8%	93.4%	94.3%	94.3%
	0.7	Bias	0.0043	0.0038	0.0028	0.0028	0.0051	0.0047	0.0055	0.0056
		SSE	0.3350	0.3488	0.2959	0.2919	0.3306	0.3577	0.2932	0.2885
		SEE	0.3208	0.3321	0.2873	0.2838	0.3270	0.3470	0.2924	0.2882
		CP	93.3%	93.2%	94.3%	94.1%	94.3%	93.3%	94.7%	94.8%
250	0.2	Bias	0.0000	0.0000	0.0000	0.0000	0.0002	0.0001	0.0003	0.0004
		SSE	0.0295	0.0302	0.0299	0.0296	0.0306	0.0323	0.0313	0.0311
		SEE	0.0291	0.0297	0.0296	0.0293	0.0304	0.0319	0.0310	0.0310
		CP	94.8%	94.7%	94.8%	94.7%	94.9%	94.7%	95.0%	95.0%
	0.5	Bias	0.0007	0.0007	0.0007	0.0007	-0.0017	-0.0018	-0.0016	-0.0017
		SSE	0.0660	0.0685	0.0611	0.0602	0.0664	0.0721	0.0612	0.0606
		SEE	0.0649	0.0672	0.0605	0.0597	0.0665	0.0712	0.0620	0.0613
		CP	94.4%	94.2%	94.4%	94.7%	94.5%	94.2%	95.3%	95.1%
	0.7	Bias	-0.0012	-0.0020	0.0010	0.0010	-0.0011	-0.0016	0.0002	0.0002
		SSE	0.2104	0.2194	0.1854	0.1831	0.2108	0.2285	0.1858	0.1829
		SEE	0.2065	0.2146	0.1828	0.1805	0.2105	0.2261	0.1862	0.1833
		CP	94.4%	94.2%	94.8%	94.6%	94.4%	94.6%	95.0%	95.2%
500	0.2	Bias	-0.0002	-0.0002	-0.0001	-0.0002	-0.0006	-0.0007	-0.0006	-0.0006
		SSE	0.0208	0.0213	0.0210	0.0207	0.0220	0.0234	0.0221	0.0223
		SEE	0.0207	0.0211	0.0210	0.0208	0.0216	0.0228	0.0219	0.0220
		CP	94.8%	95.0%	95.2%	95.3%	94.7%	94.7%	95.0%	94.9%
	0.5	Bias	-0.0005	-0.0006	-0.0003	-0.0003	-0.0004	-0.0004	-0.0007	-0.0007
		SSE	0.0470	0.0488	0.0432	0.0427	0.0466	0.0505	0.0430	0.0424
		SEE	0.0461	0.0478	0.0428	0.0423	0.0473	0.0508	0.0439	0.0434
		CP	94.8%	94.7%	94.7%	94.7%	95.1%	94.7%	95.4%	95.5%
	0.7	Bias	-0.0020	-0.0021	-0.0012	-0.0012	0.0005	0.0010	-0.0011	-0.0012
		SSE	0.1438	0.1500	0.1277	0.1259	0.1494	0.1618	0.1315	0.1300
		SEE	0.1467	0.1526	0.1295	0.1279	0.1498	0.1615	0.1320	0.1299
		CP	95.3%	95.2%	94.8%	95.0%	94.8%	95.1%	94.8%	94.8%

* CP is the coverage probability of the 95% confidence interval

Table 7: Performance of adjusted cost estimator under the exponential survival

			$P[C \leq X] = 0.3$				$P[C \leq X] = 0.4$			
n	ρ		$\hat{\beta}_I$	$\hat{\beta}_{II}$	$\hat{\beta}_{III}$	$\hat{\beta}_{IV}$	$\hat{\beta}_I$	$\hat{\beta}_{II}$	$\hat{\beta}_{III}$	$\hat{\beta}_{IV}$
100	0.2	Bias	0.0005	0.0007	0.0002	0.0004	0.0007	0.0007	0.0005	0.0004
		SSE	0.0485	0.0497	0.0493	0.0488	0.0502	0.0531	0.0514	0.0510
		SEE	0.0468	0.0476	0.0485	0.0478	0.0487	0.0505	0.0505	0.0499
		CP*	94.1%	93.8%	94.5%	94.2%	93.8%	93.0%	94.5%	94.1%
	0.5	Bias	0.0006	0.0008	-0.0002	-0.0002	0.0016	0.0014	0.0015	0.0014
		SSE	0.1076	0.1130	0.0967	0.0961	0.1089	0.1193	0.1000	0.0990
		SEE	0.1058	0.1102	0.0973	0.0965	0.1074	0.1156	0.0996	0.0987
		CP	94.0%	93.6%	94.5%	94.4%	94.3%	93.8%	94.8%	94.8%
	0.7	Bias	0.0009	0.0010	0.0023	0.0026	-0.0100	-0.0098	-0.0088	-0.0088
		SSE	0.3476	0.3666	0.2941	0.2921	0.3496	0.3881	0.2954	0.2930
		SEE	0.3367	0.3527	0.2897	0.2877	0.3399	0.3697	0.2939	0.2915
		CP	93.6%	93.4%	94.3%	94.4%	94.1%	93.3%	94.5%	94.6%
250	0.2	Bias	-0.0008	-0.0008	-0.0006	-0.0006	-0.0002	-0.0003	0.0000	0.0000
		SSE	0.0300	0.0308	0.0306	0.0302	0.0314	0.0331	0.0322	0.0319
		SEE	0.0300	0.0306	0.0307	0.0303	0.0312	0.0327	0.0320	0.0317
		CP	94.3%	94.4%	94.7%	94.5%	95.0%	94.9%	94.7%	94.8%
	0.5	Bias	-0.0008	-0.0009	-0.0002	-0.0002	0.0019	0.0020	0.0019	0.0019
		SSE	0.0681	0.0713	0.0625	0.0620	0.0703	0.0771	0.0637	0.0631
		SEE	0.0680	0.0711	0.0619	0.0614	0.0692	0.0755	0.0633	0.0628
		CP	94.9%	94.8%	94.8%	94.9%	94.5%	94.2%	94.7%	94.9%
	0.7	Bias	-0.0028	-0.0029	-0.0028	-0.0028	-0.0029	-0.0031	-0.0018	-0.0016
		SSE	0.2198	0.2310	0.1882	0.1867	0.2248	0.2503	0.1882	0.1866
		SEE	0.2173	0.2285	0.1844	0.1832	0.2202	0.2422	0.1874	0.1858
		CP	94.4%	94.6%	94.3%	94.4%	94.4%	94.0%	94.8%	94.9%
500	0.2	Bias	-0.0011	-0.0011	-0.0010	-0.0010	-0.0005	-0.0004	-0.0005	-0.0005
		SSE	0.0210	0.0216	0.0214	0.0211	0.0227	0.0240	0.0230	0.0228
		SEE	0.0213	0.0217	0.0218	0.0215	0.0221	0.0233	0.0227	0.0225
		CP	95.2%	95.3%	95.3%	95.5%	94.1%	93.6%	94.4%	94.5%
	0.5	Bias	-0.0007	-0.0008	-0.0010	-0.0010	-0.0011	-0.0010	-0.0015	-0.0014
		SSE	0.0491	0.0513	0.0447	0.0443	0.0498	0.0548	0.0449	0.0446
		SEE	0.0484	0.0506	0.0439	0.0435	0.0493	0.0539	0.0448	0.0445
		CP	94.5%	94.6%	94.9%	94.6%	94.7%	94.4%	94.7%	94.8%
	0.7	Bias	0.0027	0.0029	0.0003	0.0003	0.0000	0.0002	0.0006	0.0006
		SSE	0.1557	0.1646	0.1299	0.1291	0.1559	0.1727	0.1323	0.1311
		SEE	0.1546	0.1627	0.1307	0.1298	0.1568	0.1731	0.1328	0.1316
		CP	94.8%	94.5%	95.0%	95.1%	95.1%	94.8%	95.4%	95.4%

* CP is the coverage probability of the 95% confidence interval

Table 8: Unadjusted cost estimates stratified by controlling variables

Variables	n (%)	Average (SE)	95% confidence interval		
				Lower	Upper
Sex	Female	3243 (43.32)	340,420 (6776)	327,138	353,703
	Male	4244 (56.68)	388,329 (6454)	375,679	400,979
Age	Under 35	202 (2.69)	444,445 (42633)	360,883	528,007
	36 – 45	478 (6.38)	441,910 (21467)	399,834	483,987
	46 – 55	944 (12.6)	451,895 (15035)	422,426	481,365
	56 – 65	1542 (20.6)	398,935 (9667)	379,986	417,884
	66 – 75	2450 (32.72)	343,839 (7398)	329,338	358,341
	76 and over	1871 (25.11)	272,369 (7732)	257,214	287,525
Type of hospital	Private	5051 (67.46)	353,790 (5587)	342,839	364,742
	Public	2436 (32.54)	393,045 (8539)	376,309	409,783
Area of residence	Central	1356 (18.11)	385,320 (11632)	362,521	408,121
	Kaohsiung	1349 (18.02)	326,510 (10233)	306,452	346,569
	Northern	1280 (17.07)	312,018 (9725)	292,957	331,081
	Southern	973 (13)	335,075 (9642)	316,177	353,975
	Taipei	2529 (33.8)	422,018 (9355)	403,682	440,354
Hospital level	Regional	3817 (50.98)	374,393 (8037)	358,641	390,146
	Teaching	2441 (32.6)	227,323 (10715)	206,322	248,326
	Medical	1229 (16.42)	404,112 (6592)	391,191	417,034
Stage	I	4535 (60.57)	272,942 (4594)	263,937	281,947
	II	1256 (16.78)	509,370 (11815)	486,212	532,530
	III	683 (9.12)	712,092 (27572)	658,050	766,135
	IV	1013 (13.53)	749,552 (24541)	701,452	797,654
Treatment	Surgery & Chemo	3790 (50.62)	452,333 (6576)	439,443	465,224
	Surgery only	1510 (20.17)	290,698 (8447)	274,142	307,254
	Chemo & radio	1115 (14.89)	388,780 (14568)	360,225	417,336
	No	1072 (14.32)	83,998 (8550)	67,240	100,757
Total		7487	358,620 (4281)	349,173	368,073

Table 9: Model of fit for covariance structure

		Unweighted model			Weighed model		
	df	AIC	$-2\log L$	G^2	AIC	$-2\log L$	G^2
ANTE(1)[#1]	9	64846*	64826	–	60063*	60043	–
TOEPH[#2]	9	65066	65048	222	60359	60339	296
ARH(1)[#3]	6	65089	65075	249	60386	60372	328
CSH[#4]	6	65769	65757	931	X	X	
Independence	1	72624	72622	7796	68254	68252	8209

[#1]:ANTE(1) stands for ante-dependent covariance matrix.
[#2]:TOEPH stands for heterogeneous Toeplit covariance matrix.
[#3]:CSH stands for heterogeneous AR(1) covariance matrix.
[#4]:CSH stands for heterogeneous compound symmetry covariance matrix.
X:mod.el doesn't converge
*:model has a smaller AIC value.

Table 10: Model selections using likelihood ratio statistics

Model	Linear	Quadratic	Cubic	4th degree
Linear	–	610	2057	2073
Quadratic	–	–	1447	1463
Cubic	–	–	–	16

Table 11: Adjusted estimates

Variable	Unweighted model			Weighted model		
	$\hat{\beta}$	SE	p value	$\hat{\beta}$	SE	p value
Intercept	13.7276	0.4429	<.0001	13.8642	0.3638	<.0001
Sex						
Female vs Male	-0.0887	0.0179	<.0001	-0.0819	0.0182	<.0001
Age						
35- vs 76+	-0.0175	0.0573	0.7602	0.1236	0.0601	0.0396
36 – 45 vs 76+	0.0200	0.0398	0.6158	0.0888	0.0409	0.0299
46 – 55 vs 76+	-0.0042	0.0314	0.8943	0.0515	0.0320	0.1075
56 – 65 vs 76+	-0.0231	0.0270	0.3928	0.0195	0.0275	0.4771
66 – 75 vs 76+	-0.0685	0.0239	0.0042	-0.0442	0.0242	0.0681
Type of hospital						
Private vs public	-0.0193	0.0206	0.3475	-0.0271	0.0208	0.1942
Region						
Central vs Taipei	-0.0860	0.0261	0.0010	-0.0826	0.0265	0.0018
Kaohsiung vs Taipei	-0.2108	0.0269	<.0001	-0.2067	0.0274	<.0001
Northern vs Taipei	-0.2244	0.0277	<.0001	-0.2314	0.0281	<.0001
Southern vs Taipei	-0.0417	0.0297	0.1599	-0.0291	0.0301	0.3346
Hospital level						
Regional vs Medical	-0.1490	0.0221	<.0001	-0.1359	0.0224	<.0001
Teaching vs medical	-0.5160	0.0273	<.0001	-0.4717	0.0281	<.0001
Stage						
II vs I	0.4054	0.0252	<.0001	0.3702	0.0253	<.0001
III vs I	0.6191	0.0322	<.0001	0.5639	0.0319	<.0001
IV vs I	0.7569	0.0274	<.0001	0.7023	0.0273	<.0001
Comorbidity	0.0835	0.0066	<.0001	0.0738	0.0066	<.0001
Treatment						
Surgery & chemo vs no	3.0543	6.7400	<.0001	3.0191	0.3777	<.0001
Surgery vs no	3.7478	0.4727	<.0001	3.7789	0.4010	<.0001
Chemo & radio vs no	0.8964	1.7800	0.0751	0.8841	0.4387	0.0439
Time						
Time	-4.2170	0.6678	<.0001	-4.2025	0.5419	<.0001
Time^2	1.3468	0.2589	<.0001	1.3852	0.2056	<.0001
Time^3	-0.1383	0.0293	<.0001	-0.1450	0.0228	<.0001
Time × Treatment						
Time × surgery & chemo	-1.9857	-2.9000	0.0037	-2.2035	0.5634	<.0001
Time × surgery	-3.3441	0.7121	<.0001	-3.6238	0.5974	<.0001
Time × chemo & radio	-0.0142	-0.0200	0.9850	-0.1751	0.6557	0.7894
Time^2 × Treatment						
Time^2 × surgery & chemo	-0.0548	-1.8200	0.0680	0.6193	0.2139	0.0038
Time^2 × surgery	-0.1007	0.0311	0.0012	1.1042	0.2266	<.0000
Time^2 × chemo & radio	-0.0004	-0.0100	0.9899	0.0170	0.2504	0.9460
Time^3 × Treatment						
Time^3 × surgery & chemo	0.5528	2.0900	0.0369	-0.0609	0.0237	0.0104
Time^3 × surgery	1.0183	0.2752	0.0002	-0.1085	0.0251	<.0001
Time^3 × chemo & radio	-0.0126	-0.0400	0.9658	-0.0020	0.0279	0.9417

Acknowledgment
This research is partially supported by Nation Science Council Grant # NSC 95-2118-M-305-003.

References

Bang, H. and Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika*, 87, 329-343.

Baser, O., Gardiner, J. C., Bradley, C. J., Yuce, H., Given, C. (2006). Longitudinal analysis of censored medical cost data. *Health Economics*, 15, 513-625.

Carpenter, J. R. and Kenward, M. G. (2005). A comparision of multiple imputation and inverse probability weighting for analysis with missing data. *Journal of the Royal Statistical Society*, Series A,

Fitzmaurice, G. M., Laird, N. M., Ware, J. H. (2004). *Applied longitudinal analysis*, John Wiley and Sons, Inc., New York.

Hogan, J. W., Roy, J. and Korkontzelou, C. (2004). Tutorial in biostatistics handling drop-out in longitudinal studies. *Statistic in Medicine*, 23, 1455-1497.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assn.*, 53, 457-481.

Lin, D. Y. (2000). Linear regression analysis of censored medical costs. *Biostatistics*, 1, 35-47.

Lin, D. Y., Feuer, E. J., Etzioni, R. and Wax, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics*, 53, 419-434.

Liu, L., Wolfe, R. A. and Kalbfleish. (2007). A shared random effects model for censored medical costs and mortality. *Statistics in Medicine*, 26, 139-155.

Lo, J. C., Shih, K. S., Chen, K. L. (1996). Technical efficiency of the general hospitals in Taiwan–an application of DEA. *Taiwan Public Health J*, 15, 375-396.

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

Shih, K. S., Lo, J. C., Chen, K. L. (1996). A study on the efficiency difference between public and private general hospitals. *Taiwan Public Health J*,15, 469-482.