HEDG Working Paper 09/12

Estimating Lifetime or Episode of illness Costs

Anirban Basu
Willard G. Manning

November 2009
ISSN 1751-1976

york.ac.uk/res/herc/hedgwp

Estimating Lifetime or Episode-of-illness Costs Under Censoring

by

Anirban Basu* and Willard G. Manning**

November 14, 2009

* Section of Hospital Medicine, the Department of Medicine, and Center for Health and the Social Sciences. University of Chicago and the National Bureau of Economic Research, Cambridge, MA.
E-mail: abasu@medicine.bsd.uchicago.edu.

** Harris Graduate School of Public Policy Studies, and the Department of Health Studies of the Biological Sciences Division, University of Chicago
E-mail w-manning@uchicago.edu.

No. of Tables: 3

No of Figures: 4

Corresponding author.

Anirban Basu Ph.D.
Assistant Professor
Section of Hospital Medicine, Dept. of Medicine
The University of Chicago
5841 So. Maryland Ave, MC-2007
Chicago IL 60637

We would like to acknowledge the invaluable data management and programming support of Olena Verbenko throughout this project. We are indebted to funds from the Harris School of Public Policy Studies and the Department of Medicine in the Biological Science Division for underwriting this work. We also want to thank Paul Rathouz, Pravin Trivedi, Frank Windmejier, and seminar participants in the European Econometrics Workshop and in the Health Studies Department at the University of Chicago for their very helpful suggestions and comments.

The opinions expressed are those of the authors, and not those of the University of Chicago.

This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of the Applied Research Program, NCI; the Office of Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database.

Abstract

Many analyses of health care costs involve use of data with varying periods of observation and right censoring of cases before death or the end of the episode of illness. The prominence of observations with no expenditure for some short periods of observation and the extreme skewness typical of these data raise concerns about the robustness of estimators based on inverse probability weighting (IPW) with the survival from censoring probabilities. They also cannot distinguish between the effects of covariates on survival and intensity of utilization, which jointly determine costs. In this paper, we propose a new estimator that extends the class of two-part models to deal with random right censoring, and more fully incorporates the information from the censored periods. Our model also addresses issues about the time to death in these analyses and separates the survival effects from the intensity effects. Using simulations we highlight our proposed estimator compared to the inverse probability estimator, which shows bias when censoring is large & covariates affect survival. We find our estimator to be unbiased and also more efficient for these designs. We apply our method and compare it to the IPW method using data from the Medicare-SEER files on prostate cancer.

Keywords: Censored costs, inverse probability weighting, episode of illness, survival versus intensity effects

1. Introduction

Longitudinal profiles of heath care costs and utilization data arising out of claims databases and from clinical trials are widely used for heath technology assessments, for estimating the costs of an episode of illness and for many other purposes. Usually, the analyst can never observe all the patients until the end of their episode of illness or treatment or until they die. So an integral aspect of the analytical problem is dealing with the censoring of the observations. It has been convincingly established that traditional survival models, which are used to deal with censoring in time-to-event data, fail to address censoring in costs data even when the censoring are assumed to be random because inherent patient heterogeneity with respect to cost accumulation implies that the cumulative cost at the censoring time is positively correlated with the cumulative cost at the endpoint of interest (Lin, 1997, 2000).

In order to deal with such censoring, researchers have proposed survival-adjusted (Lin, 1997; the "LIN97" estimator henceforth) or inverse-survival probability of censoring weighted estimators that are consistent for the rate of accumulation of costs (Bang and Tsiatis 2000; Lin 2000a, 2000b; the "BTL00" estimator henceforth). Under the assumptions of continuous death and censoring times, Lin's 1997 estimator is biased (Lin, 1997). Lin avoids this problem by assuming discrete death times, which could be made to coincide with the end points for the blocks of time over which a patient's cost history is expressed. Recognizing this assumption as a limitation of the LIN97 method, Bang and Tsiatis (2000) propose an alternative estimator, based on inverse probability weighting with survival from censoring probabilities, which allows for continuous death and censoring times. This approach has also been adopted by Lin in his subsequent papers (2000a, 2000b, 2003).

Although, the BTL00 approach provides a powerful approach to deal with random censoring in cost data, there are a couple of limitations. First, an estimator following the BTL00 approach does not use information from the periods where censoring is observed.[1] This may lead to loss of efficiency in estimation, which can have serious consequences for the idiosyncratic cost distributions. Second, most estimators that apply the BTL00 methods fail to distinguish between the effect of a covariate on survival and intensity of utilization, both of which affect costs jointly. Understanding these differential effects is important giving the theoretical underpinnings of cost

[1] We consider the interval-data where an individual patient's cost-trajectory is chopped in to various time intervals. Therefore, only the interval where censoring is observed, but not any of the previous intervals, is dropped from estimation.

analyses, which are conducted to reign in inefficient utilization but to promote practices that increase survival.

In this paper, we develop a novel estimator that extends the LIN97 survival-adjusted estimator and that is consistent under the assumptions of continuous distributions of death and censoring times and also accounts for a variety of additional aspects of the cost-accumulation process, such as the widely observed increases in costs at the very end-of-life. Throughout, we assume that the analyst has data on a representative sample of the population of interest that contain some information about the rate of accumulation or timing of cost for each patient over time from the first observation (at the incidence of the illness or the initiation of the treatment episode) until the last time point at which the patient is observed. Furthermore, there are several specific issues that we address in this work that have not been addressed so far in the literature; they include – 1) use of non-linear two-part models appropriate for modeling skewed outcomes in the presence of censoring; 2) variable rates of accumulation of costs over time; 3) spikes in cost-accumulation due to end-of-life care; and 4) a more parametric approach to deal with censoring that is non-informative, which could potentially generate efficiency gains over currently existing and more popular non-parametric approaches. All our discussions consider the opportunity to adjust for covariates in estimating the mean total costs per patient. One important contribution of our estimator is that it explicitly separates the marginal effect of covariate on total costs into a portion that is brought about by affecting survival and another that is due to affecting the rates of cost-accumulation if alive.

The paper is structured as follows: Section 2 lays out a theoretical structure of cost-accumulation over the lifetime of a patient or an episode or for a fixed period of time and defines the primary parameter of interest in the presence of deaths and non-informative censoring; it describes the data that is typically available for analysis in health-care and usually arises from a long-term clinical trial or a claims database and reviews the LIN97 and BTL00 estimators. Section 2 also describes our proposed estimator. Section 3 contains simulations that illustrate the performance of our proposed estimator in terms of bias and efficiency. Section 4 illustrates the application of our estimator along with the BTL00 estimator to the analysis of prostate cancer costs. Section 5 concludes with a discussion.

2. Theoretical Model, Available Data and Estimators under Censoring

2.1. A theoretical model for accumulation of costs.

We start by laying out the theoretical model for the cost accumulation process. We assume that time begins with the start of an episode of illness or with initiation of treatment. We set the time index t for this defining event at $t = 0$ (*index time*). If the cases start at different calendar times, they are assumed to arrive at random. The time t is continuous. Our interest lies in estimating the population average of the total costs over the time period T, where at least one patient in the sample is observed to be alive until T.

We follow the notations used by O'Hagan and Stevens (2004). Let the cumulative cost up to *and including* time t for a random patient in the population be denoted by $M(t)$. $M(t)$ is a non-decreasing function of t for any specific individual. Let V denote the time to death for this patient. Formally, we suppose that $M(t) = M(V)$ for all $t \geq V$, because no further accumulation of costs occurs after death. Hence $M(T)$ can be seen as the accumulated cost to time T or death, whichever occurs earlier. Let $R(t) = \partial M(t) / \partial t$ denote the rate of accumulation of cost at anytime point t, $t \leq T$.[2] $R(t)$ is also called the (non-cumulative) cost-function. Therefore, the total costs for this patient is given by

$$M(T) = \int_{t=0}^{T} R(t)dt \text{ , where } R(t) = 0 \text{ if } t > V \qquad (1)$$

which is the area under this per-period cost-function curve until T. When time is represented as $K+1$ discrete blocks of, not necessarily equal, durations (months or year)[3] rather than continuous, let the end of each block period be denoted as $a_1, a_2, a_3, .. a_{K+1} = T$, which are the same for all patients. In this case, under some general assumptions, $R(a_j)$ can be approximated with $(M(a_j) - M(a_{j-1}))/(a_j - a_{j-1})$, $j = 1,2,.. K+1$, where $a_0 = 0$ and $M(0) = 0$. The total costs expression is then given by:

$$M(T) = \sum_{j=1}^{K+1} R(a_j) \cdot (a_j - a_{j-1}) \quad \text{where } R(a_j) = 0 \text{ if } a_{j-1} > V \qquad (2)$$

Because this is the typical format in which most observed datasets are available, we will stick to this discrete formulation for the rest of the paper.

---

[2] O'Hagan and Stevens (2004) point out that *M(t)* will typically be a step function and hence not continuous, but by definition it is always right-continuous.

[3] This approach will be useful later in practice because in many observational datasets, the last interval may be of duration that is very dissimilar to other periods of observation due to death, attrition, or administrative censoring (e.g., end of the field phase of the study or trial).

The cumulative cost *M(T)* is the principal random variable of interest, and our primary interest is in its expectation:

$$\mu = E(M(T)). \quad \textbf{(3)}$$

Because *V*, or the time of death, is stochastic in nature, one can reformulate (2) in the following two ways. Estimation following either of these empirical forms produces identical results as long as every individual in the sample is observed to the minimum (*V, T*); that is in the absence of any censoring. The first alternative form of (2) is expressed as:

$$\mu = E\left(\sum_{b=1}^{K+1} I\left(a_{b-1} < V \le a_b \right) \cdot \sum_{j=1}^{b} \left(R(a_j) \right) \cdot \left(a_j - a_{j-1} \right) \right) \quad \textbf{(4)}$$

where *I*(.) is an indicator function. Equation (4) accumulates over the individual–specific trajectories of costs and then averages over patients. This is reflective of the minimal-*data* case mentioned in Lin (1997, 2003) and in O'Hagan and Stevens (2004), where only accumulated cost per patient over a certain time interval is available. Estimation following this formulation follows a pattern mixture approach that uses fully conditional model for the outcome conditional on death times (Ribaudo et al, 2001; Pauler et al., 2003).

The second formulation of (2) can be expressed as

$$\mu = \sum_{j=1}^{K+1} Pr(V > a_{j-1}) \cdot E\left(R(a_j) \,|\, V > a_{j-1} \right) \cdot (a_j - a_{j-1}) \quad \textbf{(5)}$$

$Pr(V > a_j) = S(a_j)$ is more popularly known as the *survivor function*. Equation (5) averages period–specific costs over patients who were alive at the beginning of that period and then accumulates the costs over different periods. This representation is reflective of the *interval –data* that contain some information about the cost trajectories for each patient. Estimation using this formulation requires the estimation of a directly parameterized regression conditioning on being alive (Pepe et al, 1999; Kurland and Heagerty, 2005)

2.2. Typically available data and estimators of mean costs under censoring

In many studies, we would not expect to observe all patients until *T* or *V*. Patients whose observation period ceases before *T* or *V* are censored. Let *C* denote the censoring time. We expect to have the following format for our observed data. Each subject is observed over a minimum period of $L = \text{Min}(T, V, C)$. Time covariates *V* and *C* are continuous. Each subject provides a vector

of total costs, $\underset{\sim}{Y} = (Y_1, ..Y_b)$, over interval increments in time (say, 2 months or 6 months) where the last interval with observed costs for a subject is $(a_{b-1}, a_b]$, with $a_{b-1} < L \leq a_b$, $L = T$ for some subjects.[4] We compute a variable U_j containing the observed duration for each interval:

$$\begin{aligned} U_j &= a_j - a_{j-1} && \text{if } \min(C, V) > a_j \\ &= \min(C,V) - a_{j-1} && \text{if } a_{j-1} < min(C,V) \leq a_j \\ &= 0 && \text{otherwise} \end{aligned} \qquad \textbf{(6)}$$

Additionally, for each interval, denote a death indicator (D_j) and a censoring indicator (A_j) as follows:

$$\begin{aligned} &\text{if } \min(C, V) \leq a_j: && D_j = I(\min(C, V)=V);\ A_j = I(\min(C, V)=C); \\ &\text{if } \min(C, V) > a_j: && D_j = 0 \ \&\ A_j=0 \end{aligned} \qquad \textbf{(7)}$$

where $I(.)$ is an indicator function.

The observed data vector for each subject, where $\underset{\sim}{X}$ is a vector of covariates, is given as

$$\{\underset{\sim}{Y}, \underset{\sim}{X}, L, \underset{\sim}{D}, \underset{\sim}{A}, \underset{\sim}{U}\}$$

The data is setup so that each observation represents a subject-interval combination.

Lin (1997) proposes an estimator, LIN97, for the population unconditional mean cost in time interval j. It is given as

$$\hat{\mu}_{j,Lin} = \frac{\sum I\{Min(C,V) > a_{j-1})\} \cdot \hat{S}(a_{j-1}) \cdot \{M(a_j) - M(a_{j-1})\}}{\sum I\{Min(C,V) > a_{j-1})\}}, \qquad \textbf{(8)}$$

where the estimator adjusts the costs accumulated in the j^{th} interval by multiplying with $\hat{S}_j(a_{j-1})$, which is the Kaplan-Meier survival (from death) estimator for the (j-1)th interval, and averages over those who were alive and uncensored at the beginning of that interval. Note that the LIN97 estimator explicitly assumes that all censoring in that interval occurs at the end of that interval, i.e. at a_j. Otherwise, $\hat{\mu}_{j,Lin}$ is not a consistent estimator of μ_j because those patients observed to die in the interval (i.e. not censored) will not be a representative sample of those who actually die in that interval (Lin, 1997).

To address this limitation, Bang and Tsiatis (2000) propose a weighted complete-case estimator, BTL00, within each interval:

[4] These periods need not be of equal duration.

$$\hat{\mu}_{j,BT} = n^{-1} \sum \frac{\Delta_j \cdot (M(a_j) - M(a_{j-1}))}{\hat{K}_j \{min(V, a_j)\}}, \quad (9)$$

where $\Delta_j = I\{C \geq min(V, a_j)\}$ denotes an indicator for person intervals where the person is not censored till the end of that interval (or death, if it happens within the interval) and $\hat{K}_j(a_j)$ is the Kaplan-Meier survival (from censoring) estimator for the j^{th} interval. The rationale behind this estimator is that each subject, who we observe to die in the j^{th} interval or reach the end of the j^{th} interval without censoring, represents on average $1/\hat{K}_j(a_j)$ individuals who might have been censored. Bang and Tsiatis (2000) claim that this estimator, by focusing only on person intervals where there is no censoring, allows for censoring to occur anywhere within an interval. This inverse survival from censoring weighted estimator was also adopted by Lin (2000, 2003).

There are however some limitations to the BTL00 estimator. First, the estimator is not set up to easily separate the effect of covariates on survival versus on intensity in utilizations, both of which can be quite informative for economic analyses. Although Lin's (2003) work shows how to estimate conditional cost-estimators among patients who survive a given interval or who die in a given interval, it does not decompose covariate effects on the unconditional mean into a part that is attributable to a survival effect and a part that is due to variable rates of cost accumulation. Second, efficiency of the BTL00 estimator is called into question, a property it shares with other inverse probability weighting approaches. On one hand, when censoring is high, the estimator drops data from all intervals where a person in censored thereby loosing information and generating inefficiency. On the other hand, since the estimator requires an explicit model for time to censoring, estimators for such a censoring model could be inefficient when censoring is low, and that can add to the inefficiency of the overall estimator. To address these limitations, we develop the proposed estimator described in the next sections.

2.3. A more refined model

We propose a novel estimator that extends the LIN97 estimator to allow for continuous death and censoring times and also accounts for a variety of idiosyncratic characteristics of the cost distribution and the cost accumulation process. We follow the theoretical model in Equation (5) that is automatically set-up to handle random censoring in cost data. Both death and censoring can

occur in the middle of an interval and that should affect how we define the rate of cost accumulation function $R(a_j)$.[5] We redefine $R(a_j)$ as

$$\tilde{R}(a_j) = R(a_j) = (M(a_j) - M(a_{j-1}))/(a_j - a_{j-1}) \text{ if } \text{Min}(C, V) \geq a_{j-1}$$

$$= (M(\tilde{a}_j) - M(a_{j-1}))/(\tilde{a}_j - a_{j-1}), \text{ if } \text{Min}(C, V) > a_{j-1}, \text{ where } \tilde{a}_j = Min(a_j, C, V) \quad \textbf{(10)}$$

In intervals where death or censoring occurs, the rate is calculated using the cumulated costs within the interval up to death or censoring and divided by U_j defined earlier.

Equation (5) can be rewritten as

$$\mu = \sum_{j=1}^{K+1} Pr(V > a_{j-1}) \cdot E\left(\tilde{R}(j) \mid V > a_{j-1}\right) \cdot (a_j - a_{j-1}) \quad \textbf{(11)}$$

Note that the formulation in (11) calls for an estimator that would differ from traditional estimators of *interval-data* on costs. The observed cumulative costs $(M(j) - M(j-1))$ within an interval may not necessarily be the same as the true cumulative costs within that interval if censoring occurs in the middle of an interval. A conscious effort must be made to account for probability that the person may die within that interval but after the censoring and to assess the expected time to the death event. It is essential if we want to understand separately the utilization and survival effects; we expect many analysts to be interested in that decomposition of economic analysis. We assume that the individual rate at which we observe costs to accumulate within that interval before censoring would be the same rate at which, had there been no censoring, cost would have continued to accumulate for that individual within that interval up to the end of the interval or death, whichever occurs first.

The model in equation (11) does not require us to have censoring or death occur only at the end of intervals (as in LIN97) or to drop observations on costs during estimation for those intervals where censoring occurs in the middle of the interval (as in BTL00).[6] Instead, it allows us to

[5] This is in contrast to the model proposed by Lin(1997) where censoring times was assumed to be discrete and was assumed to occur only at the end of the intervals. Bang and Tsiatis (2000) noted that such discretization of censoring times may serve as a reasonable approximation but is not true in general. Lin (1997) mentioned that this assumption is trivial as one can redefine the time intervals so that the end of the intervals match with the censoring times. However, this could be quite complicated in practice.

[6] The assumption of constant rates of accumulation within a period is more robust when the intervals are 'thin' or of short duration. Thinner intervals lead to even more highly skewed and

parametrically extrapolate the cost accumulation beyond the censoring time for that interval period, after accounting for both the probability and time to death within that interval.

Equation (11) can be further extended to look at the different accelerated cost-accumulation towards end-of life. A large literature exists on the cost of health care at the end-of-life care and the corresponding resource utilization (Scitovsky, 1984). Recently Brown et al (2002) demonstrated the there is a U-shaped pattern of cost-history among cancer patients with the left side of the U corresponding to initial treatment and the right side reflecting a substantial spike in costs during the last 6 months of life. To incorporate this aspect of cost accumulation, we can rewrite (11)
as

$$\mu = = \sum_{j=1}^{K+1} Pr(V > a_j) \cdot \left\{ C_{ij} \cdot h(a_j) + C_{2j} \cdot (1 - h(a_j)) \right\} = \sum_{j=1}^{K+1} Pr(V > a_j) \cdot \begin{Bmatrix} E\left(\tilde{R}(j) \cdot (V - a_{j-1}) \,/\, a_{j-1} < V \le a_j \right) \cdot h(a_j) + \\ E\left(\tilde{R}(j) \,/\, V > a_j \right) \cdot (a_j - a_{j-1}) \cdot (1 - h(a_j)) \end{Bmatrix}$$

(12)

where $h(a_j) = Pr(a_{j-1} < V \le a_j \,/\, V > a_{j-1})$ is the hazard of death during interval (a_j, a_{j-1}] given that the subject was observed to be alive till a_j. Equation (12) captures the fact that the rate of accumulation of costs during the interval where the subject dies may be different from those who do not die in this interval. Therefore, $C_{1j} = E\left(\tilde{R}(j) \cdot (V - a_{j-1}) \,/\, a_{j-1} < V \le a_j \right)$ represent the expected costs of an individual if the subject dies in that interval while $C_{2j} = E\left(\tilde{R}(j) \,/\, V > a_j \right) \cdot (a_j - a_{j-1})$ represents the expected costs of the same individual had he not died in that interval. Although we have illustrated this partitioning of intervals using the one interval in which death occurs, other flexible approaches where multiple intervals leading up to death can be used to represent the different cost accumulation process preceding death. That is, one can use the expression $a_b < V \le a_{b+l} \,/\, V' > a_b$ where *l* is some fixed integer. The value of *l* can also be estimated using more semi-parametric and Bayesian estimators.

---

leptokurtotic distribution of subject-interval level costs data, as well as a larger density mass at the zero-cost level. These features support the use of two-part and other non-linear models which we utilize in our empirical section.

2.4. An estimator for the proposed model

A fundamental difference between our estimator and other proposed estimators (Lin (1997, 2000), Bang and Tsiatis (2000)) for the μ parameter is that our estimator requires a parametric extrapolation of survival and cost functions to all periods for patients after they are observed to die or be censored in the data.[7] This is typical of two-part or multi-part models used in cost and expenditure data for uncensored data (Duan et al., 1983; Blough et al., 1999; and Jones, 2000). The extrapolation of the cost-function has two parts: one for the specific time period in which censoring occurs (based on the rationale in equations (10) and (11)) and another for all time periods beyond (not including) the time period in which the censoring or death occurs.

It is important to point out that the rationale for extrapolating costs to time periods beyond observed death for a patient is to represent the population level costs for the cohort of patients with similar observed characteristics as this sampled patient, not all of whom die in that period.

Estimation follows under three parts:

a) **PART-1 :** Use a flexible accelerated failure-time model, such as those based on generalized gamma distribution for time, to estimate the individual's survival function after taking into account censoring. Let $\hat{S}_j(X)$ and $\hat{h}_j(X)$ be the estimated survivor function and the hazard function for an interval *t.* (We have suppressed the notation for individuals for clarity). The predictions are obtained for all time periods for all patients.

b) **PART-2 :** Among those subject intervals, (a_{j-1}, a_j], where we observe the subject to die, i.e. where $a_{j-1} < V \leq a_j$ & $D_j = 1$, estimate a generalized linear model (or models if a two-part specification is necessary) for the observed cost functions after conditioning on covariates X and U_j (as death can occur anywhere in the middle of the interval). Use parameter estimates from this model to predict costs, $\hat{C}_{1j}(X)$, for every subject-interval in the data. A complication arises when predicting the costs for a subject-interval where the subject is not observed to die, as if it was a *death-interval,* is that we have to account for the stochastic nature of U within that interval. That is we have to account for what would the costs be if the patient died inside that interval but at different times. To account for this, for every subject interval, we simply average the predictions that are

[7] Note that we do not extrapolate beyond the maximum time that any patient is observed in the data.

conditional of each value of *U* after weighting with the observed distribution of *U* among intervals where patients are observed to die. Therefore,

$$\hat{C}_{1j}(X) = \int \hat{C}_{1j}(X,u)\, dF(U \mid a_b < V^{obs} < a_{b+1})$$

c) **PART-3 :** Next, among those subject intervals, $(a_{j-1}, a_j]$, where patients are not observed to die including those where we only observe costs over a partial duration due to censoring, i.e. where $(D_j = 0 \& A_j = 0)$ *or* $(A_j = 1 \& a_{j-1} < C \le a_j)$, estimate a generalized linear model (or models if a two-part specification is necessary) for the observed cost functions after conditioning on covariates *X* and interval duration U_j. We use parameter estimates from this model to predict costs, $\hat{C}_{2j}(X)$, for every subject-interval in our data conditional on the full length of every interval (i.e., set $U_j = (a_j - a_{j-1})$).

d) Following (12), the estimated cost function for interval *j* for any individual is given as

$$\hat{\mu}_j(X) = \hat{S}_j(X) \cdot \left[\hat{h}_j(X) \cdot \hat{C}_{1j}(X) + (1 - \hat{h}_j(X) \cdot \hat{C}_{2j}(X) \right]$$

$$\text{and } \hat{\mu}(X) = \sum_{j=1}^{K} \hat{\mu}_j(X) \qquad \textbf{(13)}$$

We expect that by the law of large numbers, the estimator, $\hat{C}_{1j}(X)$ and $\hat{C}_{2j}(X)$ should converge in probability to $C_{1j}(X)$ and $C_{2j}(X)$ respectively. It then follows from Slutsky's theorem and the consistency of the Kaplan-Meier estimator that $\hat{\mu}_j(X)$ converges in probability to $\mu_j(X)$. We rely on simulations to establish the properties of this estimator.

The marginal effect of a covariate *X* on $\hat{\mu}(X)$ is given by

$$\frac{\partial \hat{\mu}(X)}{\partial X} = \sum_{j=1}^{K} \frac{\partial \hat{\mu}_j(X)}{\partial X}$$

$$= \sum_{j=1}^{K} \frac{\partial \hat{S}_j(X)}{\partial X} \cdot \left[\hat{h}_j(X) \cdot \left(\hat{C}_{1j}(X) \right) + (1 - \hat{h}_j(X) \cdot \left(\hat{C}_{2j}(X) \right) \right] \qquad \textbf{(14)}$$

$$+ \hat{S}_j(X) \cdot \left[\frac{\partial \hat{h}_j(X)}{\partial X} \cdot \left(\hat{C}_{1j}(X) - \hat{C}_{2j}(X) \right) + \hat{h}_j(X) \cdot \left(\frac{\partial \hat{C}_{1j}(X)}{\partial X} \right) + (1 - \hat{h}_j(X) \cdot \left(\frac{\partial \hat{C}_{2j}(X)}{\partial X} \right) \right]$$

Although the formula seems daunting, the implementation is relatively simple as one just estimates the marginal effects from each of the three regressions above and plugs them into equation (14).

One very interesting feature of this estimator is that it explicitly separates the marginal effect of each covariate on total accumulated costs into two parts: a portion that is brought about by affecting survival, and another that is due to affecting the rates of cost-accumulation. Standard error for $\hat{\mu}(X)$ and $\frac{\partial \hat{\mu}(X)}{\partial X}$ can be readily obtained via bootstrapping.

3. Simulations

3.a. Designs

Simulation 1: We start by using Lin's (2003) design points to carry out extensive simulations to evaluate our proposed estimator and to compare it to the BTL00 estimator (Bang and Tsiatis, 2000, Lin, 2000, Lin 2003) which relies on inverse probability weighting of the observed cost function using the survival probability of censoring. Following Lin (2003), the survival and censoring times are generated from the exponential distribution with mean *m* and the uniform (0,*c*) distribution respectively. Maximum follow-up time is set to 10 equally spaced intervals, (0, 10], at the end of which all survival times and cost accumulation processes are censored. We study the combinations of (*m, c*)= (5,40), (5,20), (10, 40) and (10, 20) that yield approximately 20, 30, 40 and 50% censored survival times. Additionally, we also study a (10,12) combination that was not a part of Lin's (2003) original design points, but that generates censoring of about 60% typical for many long-term administrative datasets.

Costs for individual *i* in the k^{th} interval are generated using:

$$y_{ki} = \Big[I(k=1)u_i^d + I(V_i > t_k)(\eta_i + u_{ki}) + I(t_{k-1} < V_i \le t_k)\left\{(\eta_i + u_{ki})(V_i - t_{k-1}) + u_i^f\right\}\Big] e^{\beta' X_i}, \; k = 1,.., 10; \; i = 1,..n \qquad \textbf{(15)}$$

where $\eta_i, u_{ki}, u_i^d, and\ u_i^f$ 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. As Lin (2003) describes, this data generation mechanism creates a J-shaped time patterns in costs typical of most cost accumulation processes, especially cancer patients (Brown et al, 2002). These shapes for different death cohorts are illustrated in Figure 2.

We set *X* to be a treatment indicator with 500 (= $n/2$) subjects in each of the two groups and β is set to 1. We study the coefficient on X estimated by the BTL00-estimator for comparison to

the results presented in Lin (2003). However, in order to compare results across estimators, we focus on the average incremental effect of the treatment on the cost scale rather than focusing on regression coefficients. Therefore, interest lies in the incremental effect parameter:

$$\Delta = \sum_{k=1}^{10} \left(\mu_k(X=1) - \mu_k(X=0) \right), \qquad \textbf{(16)}$$

where $\mu_k(X) = E(y_{ki} \mid X)$. We also study other variants of the design points used under Simulation 1. For each estimator and data generating mechanism, we examine 1000 replicates. Standard errors are computed from the summary statistics across the replicates. The following two simulation designs also involve 1000 replicates of n=1000 with half in the treatment group (X=1) and half not (X=0).

Simulation 2: Here we study a variant of the data generating process in (15) where the effect of *X* is allowed to differ during the end-of-life interval (slope = β +1) compared to non-end-of-life intervals (slope = β):

$$y_{ki} = \Big[I(k=1)\, u_i^d + I(V_i > t_k)\,(\eta_i + u_{ki})$$

$$+ I(t_{k-1} < V_i \le t_k) e^{X_i} \left\{ (\eta_i + u_{ki})(V_i - t_{k-1}) + u_i^f \right\} \Big] e^{\beta' X_i}, \; k = 1,.., 10; \; i = 1,..n \qquad \textbf{(17)}$$

Comparison across alternative estimators is made based on the incremental effect parameters in (16).

Simulation 3: Finally, we study another variant of Simulation 1, when the treatment *X* is allowed to affect survival as well as the costs incurred while alive. Specifically, *X* is assumed to have a 30% increase in the mean survival time, which is in line with our empirical example about the effect of lesser differentiated grades of cancer. We let survival times be generated from an exponential distribution of mean = *m**exp(0.3**X)*. Otherwise, the design points for y_{ki} are the same as in (15). Again, comparison across alternative estimators is made based on the incremental effect parameters in (16).

We apply both the BTL00 estimator and our proposed estimator to these three simulation settings. The BTL00 estimator estimates $\mu_k()$ using a log-link generalized linear model where observed costs are regressed on the *X* and indicators for time intervals weighted by the inverse probability of survival from censoring. The probability of survival from censoring is estimated using a Cox proportional hazard model accounting for death. Observations from only those

intervals where a patient is not censored for the entire interval are used for estimation. Predictions, after turning on and off the X indicator variable, from this model are made to the time intervals and the patients used in estimation, while estimates for other patient-intervals are held at zero. Predictions are averaged over all patients and summed over intervals to obtain an estimate of Δ.

The proposed estimator estimates $\mu_k()$ uses the three parts described in section 3 and expressed in equation (13). We use an exponential accelerated failure time model to estimate the probability of survival from death accounting for censoring. We use a log-link generalized linear model where observed costs are regressed on the *X*, duration till the end of observation within an interval and indicators for time intervals where patients were not observed to die (i.e. they include intervals during which censoring occurs). Third, we use another log-link generalized linear model where observed costs are regressed on the *X,* duration between start of interval and death and indicators for time intervals for those where individuals were observed to die. Predictions, after turning on and off the X indicator variable, from these models are made to all time intervals for each patient. For the third model, predictions for each patient-interval were averaged over the observed distribution of duration between start of an interval and death. Finally, predictions are averaged over all patients and summed over all intervals to obtain an estimate of Δ.

3.b. Results

Table 1 summarizes the results from our simulations. Under the Simulation 1 design, the BTL00 estimator produces unbiased estimates of the log-scale slope parameter β for *X*. This result conforms to what Lin had reported (2003). When the target of inference shifts to the cumulative costs, both our proposed estimator and the BTL00 estimator are found to be unbiased. However, the BLT00 estimator has a 4 to 11 times higher variance for the estimate of the incremental effect of X on accumulated costs, $\hat{\Delta}$, the compared to our estimator. Interestingly, the variance estimates from BTL00 decreases with increasing censoring rate, implying the important role of the survival model for time to censoring. Variance estimates from our estimator are not affected by the degree of censoring. Figure 2 presents the true levels of outcomes and the predicted means from both the estimators by time intervals and levels of *X* for one of the design datasets (m=10, c=20). We find both estimators produces unbiased estimates of time interval specific mean at both levels of *X*, although the BTL00 estimator shows slight under-prediction at higher time intervals, where censoring is high, for *X*=1. However, the differences between predictions and true values do not reach statistical significance for any of the intervals.

Under Simulation 2 design points, where treatment X also affects the differential costs of dying, the results are found to be very similar to the Simulation 1 results. The estimated standard errors from any estimator for the effect of *X* on cumulative costs are much higher under this design than the previous one, more so for BTL00 estimator than our proposed estimator. The ratio of estimated variance between BTL00 and our estimator now ranges from 4 to 17. Figure 2 presents the true levels of outcomes and the predicted means from both the estimators by time intervals and levels of *X* for one of the design datasets (m=10, c=20). The results again are very similar to that in Simulation 1, with the BTL00 showing slight under-prediction at higher time intervals for *X*=1, but none are significantly different from the true means.

Simulation 3 design points, where the treatment X has an effect on survival itself, give us different results. Here, the covariate X affects both the rate of cost accumulation and also survival. Here the efficiency gains from our estimator is not as much as it was under the previous scenarios. For the design point, (m= 18*exp(0.3*X), c=20), the bias in the overall incremental effect estimator with BTL00 does not reach statistical significance. However, when one looks at time interval-specific BTL00 estimates of conditional mean for X=1 (Figure 2), it clear that the BTL00 estimator is significantly under-predicting the mean costs for higher time points where there is more censoring. This is not the case for X=0, where survival times are shorter.

In fact, for the next design point in this simulation series, where censoring is even higher, BTL00 is found to be a biased estimator for the overall incremental effect. Our proposed estimator appears to be unbiased in this regard across all design points.

4. Empirical Example

4.a. Data

We draw data from the linked SEER-Medicare for patients with prostate cancer. SEER is an epidemiologic surveillance system consisting of population-based tumor registries designed to track cancer incidence and survival in the United States. The registries collect information about all primary cancers that a person may develop. The SEER-Medicare database consists of clinical data collected by the SEER registries linked to claims for health services collected by Medicare for its beneficiaries. The database contains information about the incident prostate cancer diagnosis and treatment provided within four months of diagnosis and includes data on the characteristics of the tumor, the demographic characteristics of the patient, and zip-code-level and census-tract-level characteristic of patient's residence at first diagnosis.

We restrict our analysis to 66-year and older male patients receiving a diagnosis of prostate cancer between 1995 and 2002, and our data run from 1994 through 2004. By including data through the end of 2004, we have follow-up data for each patient from at least 2 years up to 10 years for some patients. Each person must also have at least one year of data in Medicare before their diagnosis so that we can assess their comorbid conditions. Thus our population includes no one aged less than 66. Our data exclude outpatient pharmacy data because this study period predates the introduction of Medicare Part D.

Patients enrolled in HMOs with Medicare are dropped because Medicare does not have claims data for such participants. We also restricted patients to have eligibility for both Part A and Part B for the first two year since diagnosis because we can uniformly apply this restriction to all patients in our data. Beyond two years, loss of eligibility for either Part of Medicare is considered to be a censoring event (for example, enrollees who only have Part A coverage will have missing data for their physician services). Other restrictions include restricting the analysis to clinically localized cancer patients who received either of three treatment post diagnosis: radical prostatectomy, external-beam radiation therapy, or watchful waiting. We drop patient receiving a combination of radiation and surgery.

We identify the clinical stages for each patient at diagnosis using the definitions used by Meltzer et al (2001). According to their definitions, clinical stage A comprise of tumors that are clinically localized and non-palpable on rectal exams; B, clinically localized but palpable and C, palpable with evidence of local extension beyond the prostate. Cancer grade was classified as Well, Moderate of Poor based on the Gleason Score.

The cost trajectories of patients since diagnosis are calculated with two-month intervals until death, censoring or the end of the 10 year period. By construction, every patient is allowed to possess 60 two-month intervals in the dataset corresponding to a 10 year period. However, observed costs were missing for intervals beyond the death or censoring interval for a patient. We treat all costs after death as true zeroes, while costs between censoring and death are treated as missing. Our primary outcome was cumulative medical expenditures over 10 years. In all our analysis, we adjust for year of diagnosis, demographics such as age, race, marital status, Charlson co-morbidity index, and indicators for the Elixhauser co-morbid conditions both in the year preceding diagnosis, health care expenditure quartiles in the year prior to diagnosis and zip-code level characteristics ((income, education, and racial mix). See Table 2 for a list of the covariates and their sample means.

Formally, the criterion for having prostate cancer is having an ICD-O-3[8] diagnosis indicating prostate cancer. 146,174 cases were provided to us by SEER / Medicare. Our exclusions on coverage and stage of illness lead to a drop of 37 percent. This leaves us with a potential sample of 63 percent (92,494 cases). Our estimation uses a random 10 percent of these cases (N = 9,250).

Expenditures are for a two month period or any fraction thereof if the patient died or is censored during the two month period. All inpatient (facility) charges are assigned to the period containing the admission date. No attempt was made to spread these costs over the time spanned by the admission or discharge date. All back-to-back hospitalizations are treated as one hospital stay. All expenditures are Medicare allowed charges. Adjustments have been applied. Expenditures are in nominal terms.

4.b. Estimators

We apply both a BTL00 estimator and also our proposed estimator to this data. For the BTL00 estimator, we use a two-part model to model the zero part (with a logistic model) and the non-zero part (log-link Gamma GLM model). The adjusted Kaplan Meier estimator for surviving censoring is obtained using a Cox proportional hazards model. These probability estimates are used to inverse weight the mean predictions from the two-part model.

In addition, we also apply our proposed estimator this data. Our estimator comprises of three parts as follows:

ESTM1: An accelerated failure time model with generalized gamma regression to estimate probability of survival from death and hazard of dying in any interval.

ESTM2: A Gamma GLM model with square-root link to estimate the cost function for intervals where death was observed

ESTM3: A two-part model (logistic for the zero part, and an EEE model (Basu and Rathouz, 2005) for the non-zero part) to estimate the cost function for intervals where death was not observed.

For the survival estimators for censoring (as in BTL00) or death (as in our model), covariate list includes those listed in Table 2. For each of the cost estimators, in addition to these covariates in

[8] International Classification of Diseases for Oncology, Third Edition.

the specification for the linear predictor a continuous measure of time in the form of the 2-month interval number since diagnosis, year-since-diagnosis indicators and also interaction of the year dummies with time are also included. ESTM2 additionally contains the duration variable (eq (6)) as a covariate.

We use the two-part model for the BTL00 and the ESTM3 part of our estimator to enhance the robustness of the estimation. For periods of observation less than a year, here two months, medical expenditures have the characteristic that a substantial fraction of the cases will not have any expenditure in the two-month period. For the reminder, expenditures are quite skewed, especially if there is an inpatient stay. We do not use least squares regression on log expenditures for any of the estimators because of the retransformation problem in the presence of heteroscedasticity in the cost data (Manning, 1998; Mullahy 1998). The two-part models are more robust for health expenditure data for fixed periods (typically a year).

We perform a variety of goodness of fit tests for both estimators using costs from the observed patient-intervals with no death or censoring. We present time-interval specific and cumulative results on grade-specific costs for both the 2-year since diagnosis mark (since there was no censoring till two years) and also the 10 year -since diagnosis mark. Standard errors are obtained via 500 clustered bootstrap replicates.

4.c. Results

Descriptive statistics for various patient characteristics, stratified by grade at diagnosis are given in Table 2. Patient diagnosed with well and moderate grade are slightly younger than those diagnosed with poor grade. They also have slightly lower expenditures in the pre-period. Many of the other comorbidities appear to be moderately balanced across grades.

Both the BTL00 and also our estimator pass many of the goodness of fit tests, with Pearson correlations between raw-scale residuals and predictions of -0.03 and -0.005, respectively, and no systematic patterns in residuals across deciles of the predictions.

The adjusted survival (from death) graph by grade is displayed in Figure 3(a). In line with clinical knowledge, we find that better grade at diagnosis is associated with better survival. Compared to poor grade, patients diagnosed with well grade have a 33% (95% CI: 20%, 46%) and patients diagnosed with moderate grade have a 30% (95% CI: 23% to 37%) larger mean time to death. Survival times for patient diagnosed with well and moderate grades are not significantly

different from each other. Figure 3(b) illustrates the probability of "survival from censoring" in our dataset, which is primarily used to compute the inverse probability weight for the BTL00 estimator. Figures 3 (c) and (d) illustrate the grade-specific estimated trajectory of costs under our proposed estimator and the BTL estimator respectively. The biggest discrepancy between these two estimators comes during the later intervals, where censoring is considerably high.

Our proposed estimator shows a flatter trend in costs from fourth year onwards compared to BTL00 estimator, with the costs of patient diagnosed with well or moderate grade cancer reaching the levels of those diagnosed with poor grade from the seventh year onwards. We attribute this to the better survival associated with well or moderate grade cancer compared to poor grade cancer. We do not see such trends with the BTL00 estimator. In fact, the interval specific differences in estimated mean between BTL00 and our estimator are statistically significant from the beginning of the third year for well and moderate grades and from the beginning of the seventh year for poor grade cancers.

Table 3 presents the 2-year and 10-year grade-specific results. For the 2-year-since-diagnosis mark, we do not see substantial difference (differences were statistically significant for poor grade) between our estimator and the BLT00 estimator, although our estimator is slightly more efficient. In fact estimates from our estimator are found to be more in line with the poor grade-specific cumulative observed costs in this time period of $18,036. Both our proposed estimator and the BTL00 estimator find significant differences in 2-year costs between moderate and poor and also between well and poor grades; however, these differences were significantly higher for BTL00 than ours. More interestingly, our estimator finds much of the difference in cumulative costs in the first two years between grades is due to intensity of utilization rather than due to survival effects, which is in line with the clinical knowledge in this area.

Since there are no censoring in this data during the first two years after diagnosis, the discrepancies in estimates for poor grade point out to the challenges in modeling a heterogeneous cost-accumulation process without accounting for the different parts of the distribution moving at different paces, like what we have attempted to capture with our estimator. In fact, to further check the overall specifications of either BTL00 or our estimator, we compare the predictions from both the estimators as well as to the observed costs in the first two years (Figure 4). Any differences in this period can be attributed to specification problems alone. We find that much of the difference between estimators in the first two years is due to BTL00 over-predicting the costs on the first

interval; but starting from the second interval there are no major systematic patterns of bias for either estimator .

Next, we compared the grade-specific 10-years cumulative costs estimated by BTL00 and our proposed estimator (Table 3). We find that compared to our estimator, the BTL00 estimates for the 10-year cumulative costs produces lower estimates of cumulative costs for all grades, and especially so for well and moderate grade that have better survival, which is in line with our simulation results. Our estimates of the incremental costs between poor grade and well or moderate grades are half as those of the BTL00 estimator, the differences being statistically significant. Our estimator also finds that if we only look at the increased intensity of utilization for patients diagnosed with poor grade, the incremental costs would have been close to what BTL00 estimates. However, the fact that poor grade is associated with decreased survival induces cost-savings that are highly significant. Accounting for these effects brings down the overall incremental effects of poor grade on costs compared to well/moderate grades.

We do not find major differences in efficiency between the two estimators in this empirical example. However, such a comparison may be unfair as with costs data the variance is often a power function of the mean, and the BTL00 estimator produces lower (potentially biased) estimates of the mean.

5. Conclusions.

In this paper, we have presented a method for dealing with cost data that are censored at random, such as the type that occurs if a study ends before all the patients have died or their episode of either illness or treatment ends. This is a common occurrence in studies that look at individuals with chronic illnesses or illnesses that can affect the survival of patients. Our model builds on the prior literature on censored cost analysis, particularly the work by Bang and Tsiatis (2000) and Lin (various dates, esp. 2000 and 2003) that relies on inverse probability weighting (IPW). In this work, we have subjected these earlier models and our own to comparisons using the simulation design proposed by Lin (2003) augmented by some additional data generating designs that involve situations where individuals can die within, rather than at the end of a fixed period of observation, and where the treatment variables can also affect the likelihood of surviving.

What we find is that existing estimators based on inverse probability weighting to address censoring can be sometimes biased (especially when censoring is large and covariates affect

survival) and inefficient estimators for the incremental effect of treatment on costs or Average Treatment Effects assessed on raw or actual cost scale. This last result (inefficiency) is a well known property of IPW approaches.

Our new estimator addresses this type of data with both censoring and deaths. At least for the simulations that Lin (2003) and we have considered, our new approach appears to be consistent under simulation. We plan to use a more extensive set of simulations to see if this property holds up.

In addition to the simulation work, we have applied the Bang-Tsiatis-Lin IPW (BTL00) approach and our proposed estimator to data on prostate cancer survival and Medicare costs and estimate incremental effect of cancer grade at diagnosis on costs. The results indicate that their approach and ours generate similar results when looking a 2-year cost where there is no censoring and the differential effects of grade on survival are small. However, when looking at the 10-year costs, the estimated effects of grade of the cancer differ substantially across estimators. This appears to be the result of the differential effect of grade on survival over the longer period, which conforms to the Simulation 3 result of bias in the BTL00 approach when there is an effect of treatment or other covariates on survival but those effects are partly masked due to heavy censoring.

Future work developing the proposed model further and validating it with other datasets will be quite useful.

Table 1: Simulation results for the estimation of β and Δ using BTL and our proposed estimators

% Censoring	m	c	β under BTL00-estimator		Δ under BTL00-estimator		Δ under our estimator		Var(BTL00) /Var(Our Estimator)
			Bias	SE	Bias	SE	Bias	SE	
Simulation 1									
20	5	40	0.0009	0.061	-0.25	2.61	-0.007	0.78	11.2
30	5	20	-0.0005	0.078	1.22	2.05	-0.03	0.83	6.1
40	10	40	-0.0026	0.040	-0.39	2.05	0.02	0.70	8.6
50	10	20	0.0004	0.048	-0.90	1.73	0.11	0.76	5.2
60	10	12	-0.0012	0.072	-2.32	1.52	0.03	0.79	3.7
Simulation 2									
20	5	40	-	-	2.28	5.79	-0.61	1.41	16.9
30	5	20	-	-	0.48	4.38	-0.59	1.51	8.4
40	10	40	-	-	1.34	3.98	-0.42	1.25	10.1
50	10	20	-	-	0.21	3.12	-0.49	1.32	5.6
60	10	12	-	-	-1.89	2.99	-0.49	1.42	4.4
Simulation 3									
-	5*exp(0.3X)	40	-	-	-0.50	2.16	1.20	0.80	7.3
-	5*exp(0.3X)	20	-	-	-1.20	1.70	1.22	0.84	4.1
-	10*exp(0.3X)	40	-	-	-1.22	1.75	1.25	0.70	6.25
-	10*exp(0.3X)	20	-	-	-1.81	1.58	1.25	0.70	5.1
-	10*exp(0.3X)	12	-	-	**-3.40**	1.29	1.29	0.97	1.8

Bias averaged over 1000 replicates each of n=1000. SE = standard deviation of estimates across 1000 replicates. Var(.) = SE^2.
Simulation 1: True value of Δ under the (m, c) = (5,40) and (5,20) mechanisms is 19.9 and under the (10, 40), (10, 20) and (10, 12) mechanisms is 21.1.
Simulation 2: True value of Δ under the (m, c)= (5,40) and (5,20) mechanisms is 44.15 and under the (10, 40), (10, 20) and (10, 12) mechanisms is 38.85.
Simulation 3: True value of Δ under the (m, c)= (5*exp(0.3X),40) and (5*exp(0.3X),20) mechanisms is 20.3 and under the (10*exp(0.3X), 40), (10*exp(0.3X), 20) and (10*exp(0.3X), 12) mechanisms is 21.1.

Bold-face indicates significant bias at 5% level.

Table 2 Dependent variables and covariates: Labels and sample means.

Variables (% or Mean(sd))	Grade Well (N=614)	Grade Moderate (N=6805)	Grade Poor (N=1831)	Overall (N=9250)
Pre-period expenditures quintiles**				
Quintile 1	27.0%	29.5%	25.9%	28.6%
Quintile 2	11.1%	11.5%	13.2%	11.8%
Quintile 3	21.5%	19.5%	21.2%	20.0%
Quintile 4	18.2%	20.9%	20.4%	20.6%
Zip-code level characteristics (2000 Census)				
% Non High School grads	19.3	17.7	18.5	18.0
% High School only	25.2	25.0	25.9	25.2
% Some college	28.7	28.4	28.4	28.4
% At least 4 years College	26.7	28.8	27.2	28.3
% Blacks	8.6	10.1	10.6	10.1
% Whites	69.7	71.8	69.8	71.3
% Hispanics	16.6	14.1	13.9	14.2
Pre-period comorbidities				
Valvular disease	6.2%	5.3%	5.2%	5.4%
Pulmonary circulation disease	0.2%	0.6%	0.4%	0.5%
Peripheral vascular disease	7.3%	5.7%	6.0%	5.9%
Hypertension	32.1%	31.2%	31.5%	31.3%
Paralysis	1.8%	0.9%	0.8%	1.0%
Neurological disorders	4.1%	2.2%	2.7%	2.4%
Chronic pulmonary disease	11.9%	9.7%	9.8%	9.9%
Diabetes w/ Chrn. Comp.	3.3%	2.2%	2.3%	2.3%
Hypothyroidism	5.0%	4.9%	4.8%	4.9%
Renal failure	2.1%	1.6%	1.6%	1.6%
Liver disease	0.3%	0.4%	0.7%	0.4%
Solid tumor w/o metastasis	5.2%	4.6%	5.6%	4.8%
Rheumatoid arthritis	1.5%	1.8%	1.1%	1.6%
Coagulopthy	1.1%	1.3%	1.6%	1.4%
Obesity	0.2%	0.9%	1.3%	0.9%
Weight loss	0.3%	0.8%	0.3%	0.7%
Electrolyte disorder	4.6%	4.0%	4.0%	4.1%
Chronic blood loss anemia	1.1%	0.7%	0.3%	0.7%
Deficiency anemias	8.5%	7.7%	8.5%	7.9%
Alcohol abuse	0.3%	0.6%	0.5%	0.6%
Psychoses	0.8%	0.7%	0.9%	0.7%
Depression	2.1%	1.6%	1.6%	1.7%
Other demographics & clinical conditions at diagnosis				
Age	74.8 (5.8)	74.2 (5.4)	76.2 (6.4)	74.6 (5.7)
White	79.6%	82.9%	79.4%	82.0%
Black	8.8%	10.0%	11.3%	10.1%
Single	7.7%	6.8%	6.7%	6.9%
Married	67.9%	70.7%	67.7%	69.9%
Stage1 cancer	0.0%	54.2%	42.2%	48.2%
Stage2 cancer	7.2%	12.4%	18.0%	13.2%

** Quintiles calculated based on total expenditure of entire sample, before any exclusion was applied.

Table 3: Comparison of estimated 10 year costs by grade. Means and (Standard errors)

	BTL00 estimator		Proposed estimator			
	2-year Costs	Incremental Costs	2-year Costs	Total Incremental Costs	Incremental Costs due to Rate†	Incremental Costs due to Survival††
Well	14,594 (762)+	-	15,009 (595)	-		
Moderate	15,408 (243)+	414 (803)	15,719 (234)	710 (603)	727 (598)	-16 (48)
Poor	19,178 (603)+	4,310 (606)+	18,365 (491)+ *	2,646 (435)+ *	2,870 (438)+	-225 (50)+
	Poor vs. Well→	4,724 (886)+	Poor vs. Well→	3,356 (688)+ *	3,600 (703)+	-244 (65)+
	BTL00 estimator		Proposed estimator			
	10-year Costs	Incremental Costs	10-year Costs	Total Incremental Costs	Incremental Costs due to Rate*	Incremental Costs due to Survival**
Well	40,267 (5,462)+	-	54,248 (7,055) *	-		
Moderate	43,015 (5,416)+	2,748 (2,405)	56,607 (6,919)+ *	2,359 (2,796)	2,927 (2390)	-569 (1481)
Poor	54,163 (6,962)+	11,148 (2,228)+	61,602 (7,643)+ *	4,995 (1,972)+ *	11,986 (2304)+	-6,991 (1574)+
	Poor vs. Well→	13,896 (3,172)+	Poor vs. Well→	7,354 (2,928)+ *	15,019 (3407)+	-7,666 (2088)+

Note: + indicates significant at the 5 percent level or better.
* indicates estimate significantly different from corresponding BTL00 estimate at the 5% level.
† Holding survival constant at that of the less advanced grade.
†† Holding Rate of cost accumulation constant at that of the advanced grade.

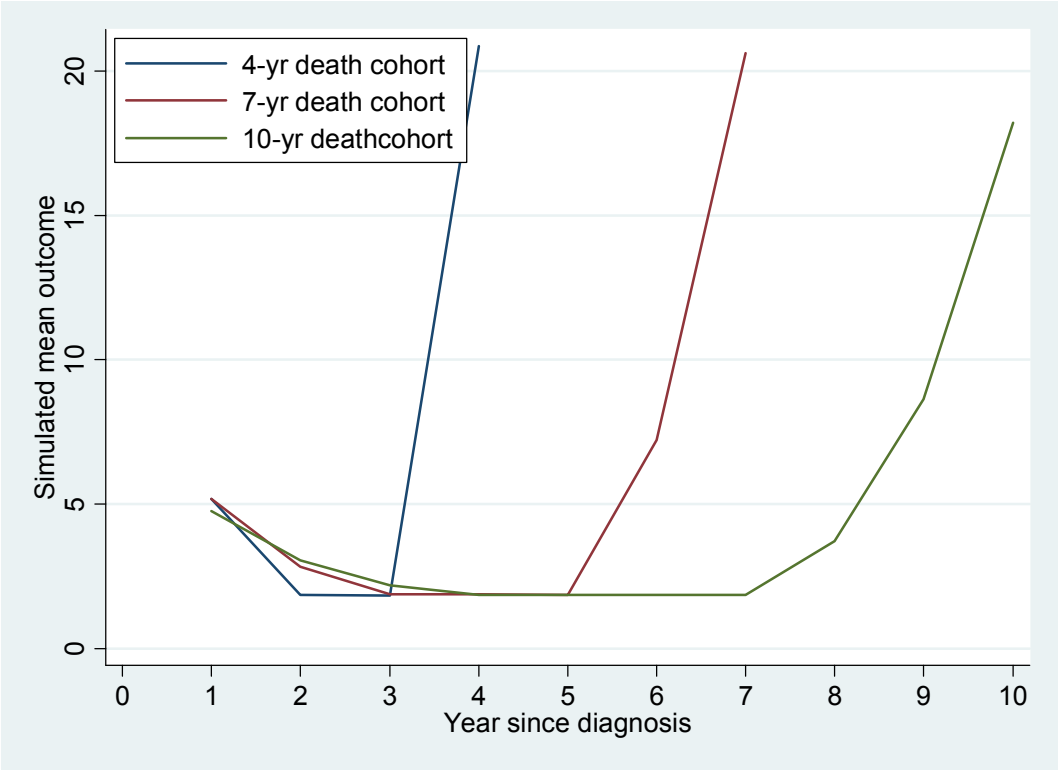

Figure 1: The J-shaped cost structure for our Simulation 1 data.

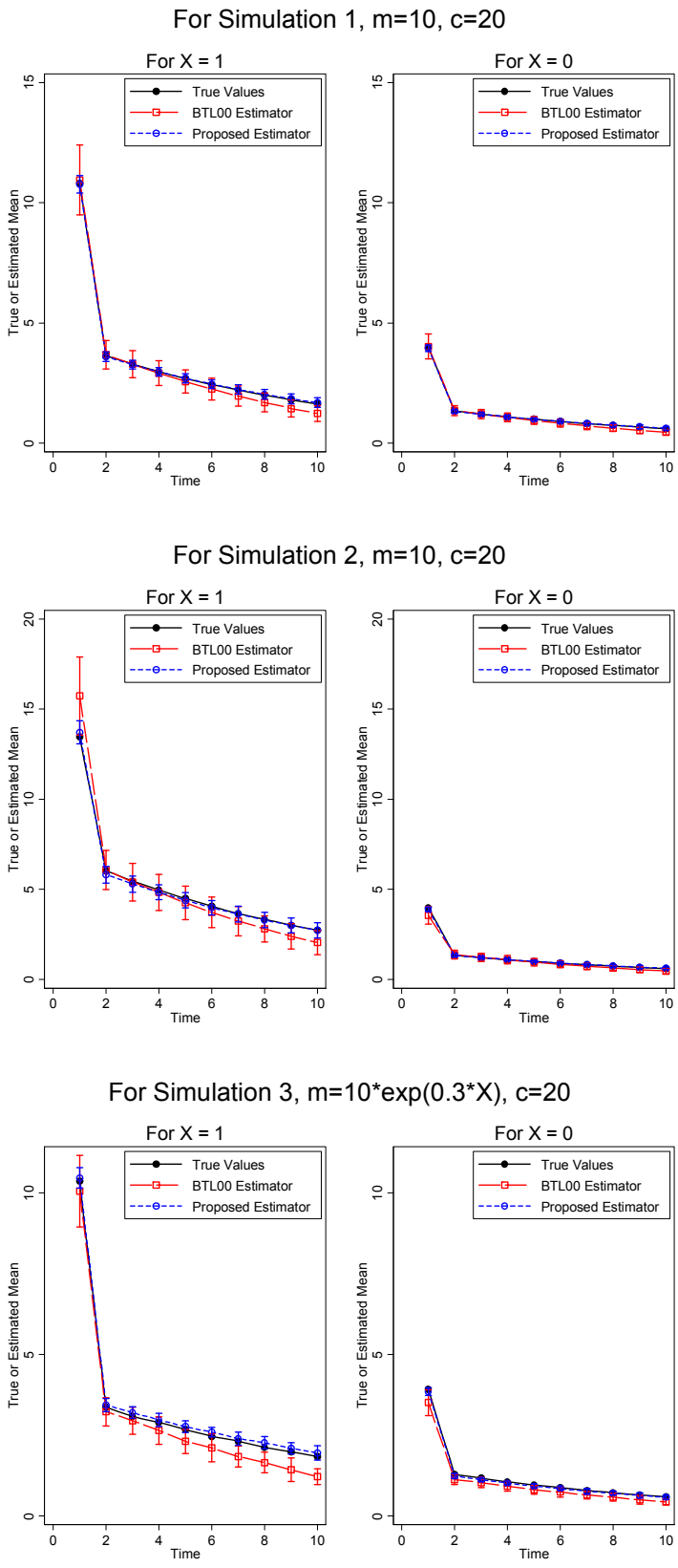

Figure 2: Estimated conditional (on X) time profiles using BTL00 and proposed estimator compared to true values.

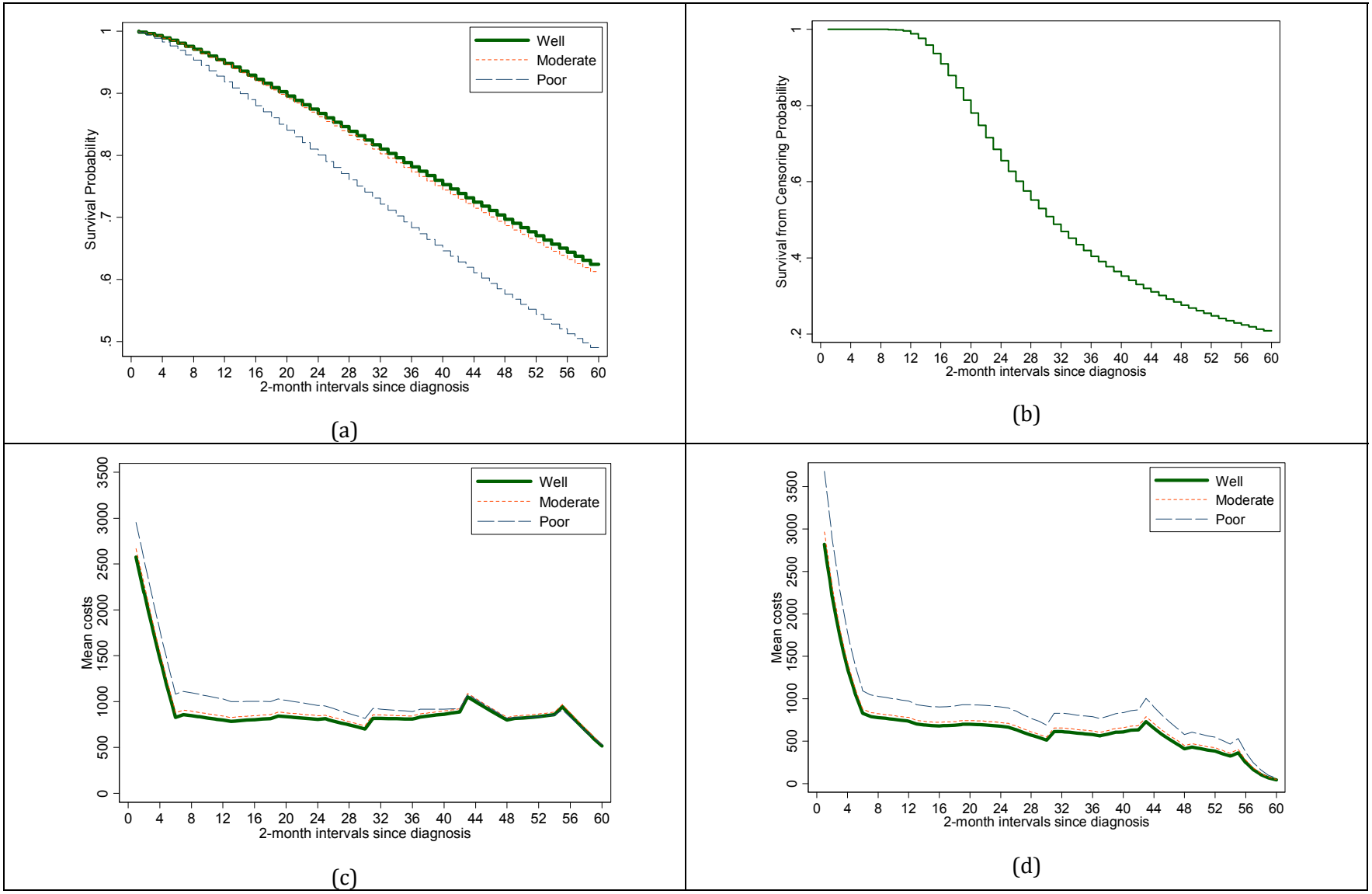

Figure 3: (a) Predicted survival (from death) by grade; (b) predicted survival (from censoring); (c) mean cost profiles by grade predicted using proposed estimator; (d) mean cost profiles by grade predicted using BTL00 estimator.

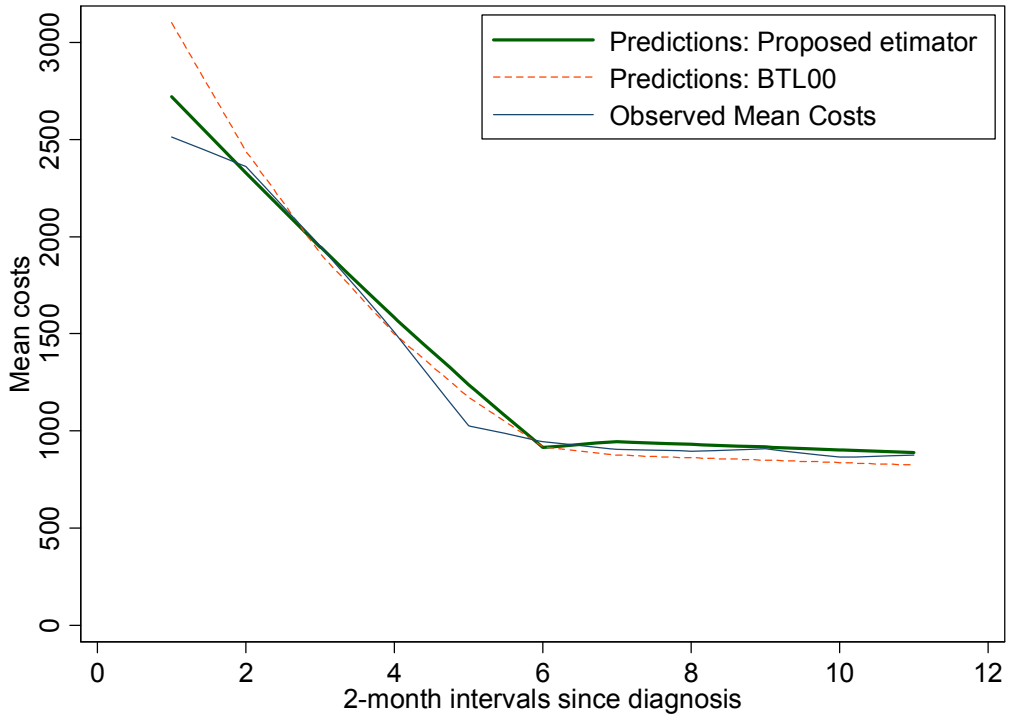

Figure 4: Goodness of fit for proposed and the BTL00 estimators during the first 22 months after diagnosis, during which there was no censoring by construction of the dataset.

REFERENCES

Bang, H., and A.S. Tsiatis. Estimating Medical Costs with Censored Data. Biometrika. 2000. **87(2)**; 329-343.

Basu, A., W.G. Manning, and J. Mullahy, "Comparing Alternative Models: Log vs. Cox Proportional Hazard?" *Health Economics*, 2004; **13(8)**: 749-765.

Blough, D.K., C.W. Madden, and M.C. Hornbrook. "Modeling risk using generalized linear models," *Journal of Health Economics*, 1999; **18:** 153-171.

Brown, M., G.F.Riley, N. Schussler, and R. Etzioni. Estimating health care costs related to cancer treatment from SEER-Medicare data. *Medical Care.* 2002; **40(8 Suppl.)**:IV-104-117.

Duan, N., W.G. Manning, C.N. Morris, and JP/ Newhousel. "A Comparison of Alternative Models for the Demand for Medical Care," Journal *of Business and Economics Statistics*,1983; **1**:115-126.

Dudley RA, Harrell Jr. FE, Smith LR. *et al*. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *Journal of Clinical Epidemiology* 1993; **46(3):** 261-271.

Fenn, P, McGuire A, Phillips V. *et al*. The analysis of censored treatment cost data in economic evaluation. *Medical Care* 1995; **33(8):** 851-863.

Etzioni RD, Feuer EJ, Sullivan SD, e*t al.* On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics* 1999; **18:** 365-380.

Hallstrom A, Sullivan SD. On estimating costs for economic evaluation in failure time studies. *Medical Care* 1998; **36(3):** 433-436.

Jain, A.K., and R.L. Strawderman. Flexible hazard regression modeling for medical cost data. *Biostatistics*, 2002, **3**:101-118.

Jones, A. "*Health Econometrics*," in A. Culyer and J. Newhouse, (Eds.), Handbook of Health Economics. Amsterdam: Elsevier, 2000.

Kurland, B. F. and Heagerty, P. J. Directly parameterized regression conditioning o being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 2005, **6(2)**:241-258.

Lin, D.Y., Feuer, E.J., Etzioni, R. and Wax, Y., 1997. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997; **53:** 113–128.

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

Lin, D.Y. Proportional means regression for censored medical costs. *Biometrics*. 2000; **56(3)**:775-8.

Lin, D.Y. Regression analysis of incomplete medical cost data. *Statistics in Medicine*, 2003; **22(7)**:1181-200.

Lipscomb J, Ancukiewicz M, Parmigiani G., *et al*. Predicting the cost of illness: A comparison of alternative models applied to stroke. *Medical Decision Making* 1998; **18(2):** S39-S56.

Manning, W.G. "The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem," *Journal of Health Economics.* 1998; 17: 283-295.

Manning, W.G., A. Basu, and J. Mullahy "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data," *Journal of Health Economics* 2005, **24**: 465-488.

Meltzer, DO, B Egleston, I Abdalla. Patterns of prostate cancer treatment by clinical stage and age in the United States. *American Journal of Public Health* 2001; 91(1): 126-128.

Mullahy, J. "Much Ado about Two: Reconsidering Retransformation and the Two-part Model in Health Econometrics," *Journal of Health Economics.* 1998;17: 247-281.O'Hagan, A., and J.W. Stevens. On Estimators of Medical Costs with Censored Data. *Journal of Health Economics*. 2004; **23:** 615-625.

Pauler, D.K., McCoy, S., and C. Moinpour. Pattern mixture models for longitudinal quality of life studies in advanced stage diseases. *Statistics in Medicine* 2003; 22: 795-809.

Pepe, M. S., Heagerty, P. and R. Whitaker. Prediction using partly conditional time-varying coefficients regression models. *Biometrics* 1999, 55: 944-950.

Raikou, M., and A. McGuire. Estimating Medical Costs Under Conditions of Censoring. *Journal of Health Economics* 2004; **23**: 443-470.

Ribaudo, H.J., Thomson, S.J., and T.G. Allen-Mersh. A joint analysis of quality of life and survival using a random effect selection model. *Statistics in Medicine* 2001; 20: 1173-1184.

Scitovsky, A.A. The high cost of dying': what do the data show? *Milbank Memorial Fund Quarterly*. 1984; 62(4):591-608.

Willan, A.R., D.Y, Lin, and A. Manca. Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine*. 2005; **24**:131-145.