Biostatistics (2015), **0**, 0, *pp.* 1–23
doi:10.1093/biostatistics/Median*Cost*Biostatistics

Median Cost Analysis Associated with Recurrent Episodic Illnesses in Presence of Terminal Events

RAJESHWARI SUNDARAM[1*], LING MA[1], and SUBHASHIS GHOSHAL[2]

[1]*Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, Rockville, Maryland 20852, U.S.A.*
[2]*Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.*

sundaramr2@mail.nih.gov

Summary

Recurrent events are often encountered in medical follow up studies. In addition, such recurrences have other quantities associated with them that are of considerable interest, for instance medical costs of the repeated hospitalizations and tumor size in cancer recurrences. These processes can be viewed as point processes, i.e. processes with arbitrary positive jump at each recurrence. An analysis of the mean function for such point processes have been proposed in the literature. However, such point processes are often skewed, leading to median as a more appropriate measure than the mean. Furthermore, the analysis of recurrent event data is often complicated by the presence of death. We propose a semiparametric model for assessing the effect of covariates on the quantiles of the point processes. We investigate both the finite sample as well as the large sample properties of the proposed estimators. We conclude with a real data analysis of the medical cost associated with the treatment of ovarian cancer.

Key words: Recurrent events; Quantile regression; Informative censoring; estimating equations; Survival

*To whom correspondence should be addressed.

analysis.

1. Introduction

Recurrent event data frequently arise in medical and epidemiological studies, where each subject may experience a number of "failures" over the course of follow up. Examples of recurrent failure events include tumor recurrences, recurrent infections, and repeated hospitalizations. In most of these examples, an additional process denoting the size of the tumor or cost associated with treating infection or hospitalization, is also of interest. These processes are no more simple recurrent events with jumps of 0 or 1 but are point processes with arbitrary positive jumps. An additional challenge in such studies is that the follow-up time is also informatively censored by terminal events, like death. Statistical methods for analyzing recurrent events has received considerable attention starting with the works of [Prentice *and others* (1981); Andersen and Gill (1982); Pepe and Cai (1993); Lin *and others* (2000); Wang *and others* (2001)]. Approaches to analyze recurrent events data in presence of terminal events based on correlated marginal modeling approach to model these two processes was proposed by Ghosh and Lin (2003) where they used a partially specified correlation structure for the dependence of the two processes. Huang and Wang (2004) proposed a joint modeling approach where the dependence between the recurrent times and failure times was captured through a frailty term. However, very limited work has been done in dealing with point processes. Strawderman (2000) proposed nonparametric estimate for the mean of a point process. Lin *and others* (2001) proposed a semiparametric transformation model for the point processes based on transformation models and Sundaram (2007) proposed modeling in presence of terminal events.

Motivated by medical costs data to assess the cost over survival time, methods for assessing mean cost using a linear regression was proposed by Lin (2000) as well as proportional means model by Lin (2000). However, medical costs are usually skewed to the right and may have

excess of near-zero costs. Consequently, quantiles may be more appropriate for assessing costs. Motivated by these issues, Huang (2002) proposed a calibration regression model for censored lifetime medical costs. Quantile regression for censored medical cost has been studied by Bang and Tsiatis (2002). Recently, quantile regression for the gap times associated with recurrent events has been studied by Luo *and others* (2013). However, assessing cost associated with recurrent hospitalizations have not been studied, especially using quantile regression. We propose a joint modeling approach for assessing the quantiles of the cost process associated with a recurrent event process with a frailty that captures the dependence between the recurrent event process and the survival times. We further model the cost associated with each recurrent event using a gamma distribution.

The outline of the paper is as follows. In Section 2, we present our notation and the proposed modeling approach for complete data as well as for data in presence of a terminal event. We also discuss the large sample properties in Section 3. The finite sample properties are investigated through simulations in Section 4 and an analysis of the cost associated with ovarian cancer patients from the SEER-Medicare medical costs is presented in Section 5. In this paper, we considered the incident part of the database.

2. Notation, Semiparametric Model

2.1 *Notation*

Let $\{N(t) : 0 \leqslant t \leqslant \tau, \}$ denote the observed counting process for an individual up to time τ, which denotes the end of the study. At this stage, we assume it to be fixed. The event times, i.e., the times of occurrence of the recurrent events are denoted by $0 < t_1 < t_2 < \cdots t_m < \tau$. We assume that the recurrent event process $N(\cdot)$ follows a non-homogeneous Poisson process with intensity/rate function $\lambda_0(t)e^{U'\gamma}$, where U is a vector of covariates and γ the corresponding vector of regression coefficients and $\lambda_0(\cdot)$ denotes the baseline intensity function. The cumulative

rate function $\Lambda(t)$ is given by

$$\Lambda(t) = \Lambda_0(t)e^{U'\gamma} = \int_0^t \lambda_0(s)e^{U'\gamma}ds.$$

A medical cost ω_j is associated to each event time t_j. We will assume for now that ω_j are i.i.d. Gamma(α, β) and moreover, ω_j are independent of $N(\cdot)$. Recall that the gamma density for the cost ω_j is given by

$$f(\omega) = \frac{\beta^\alpha}{\Gamma(\alpha)}e^{-\beta\omega}\omega^{\alpha-1}, \text{ where } \omega > 0 \text{ and } \Gamma(\alpha) = \int x^{\alpha-1}e^{-x}\,dx.$$

Thus, the total cost $\mathcal{S}(t) = \sum_{j=1}^{N(t)} \omega_j$ given $N(t) = k$ follows Gamma$(\alpha k, \beta)$.

2.2 *Modeling in presence of independent right censored data*

Note that $\mathcal{S}(t) = 0$ if and only if $N(t) = 0$. Consequently, for $x \geqslant 0$, we have

$$\begin{aligned}
&P(\mathcal{S}(t) \leqslant x|U) = \\
&P(N(t) = 0|U) + \sum_{k=1}^{\infty} e^{-\Lambda_0(t)e^{U'\gamma}} \frac{(\Lambda_0(t)e^{U'\gamma})^k}{k!} \frac{\beta^{\alpha k}}{\Gamma(\alpha k)} \int_0^x e^{-\beta y} y^{\alpha k-1}\,dy \\
&= e^{-\Lambda_0(t)e^{U'\gamma}} + \sum_{k=1}^{\infty} e^{-\Lambda_0(t)e^{U'\gamma}} \frac{(\Lambda_0(t)e^{U'\gamma})^k}{k!} \frac{1}{\Gamma(\alpha k)} \int_0^{\beta x} e^{-y} y^{\alpha k-1}\,dy.
\end{aligned}$$

Using the convention that for $k = 0$,

$$\frac{1}{\Gamma(\alpha k)} \int_0^x e^{-\beta y} y^{\alpha k-1}\,dy = 1 \text{ for any } x \geqslant 0,$$

one can write

$$P(\mathcal{S}(t) \leqslant \frac{x}{\beta}|U) = \sum_{k=0}^{\infty} e^{-\Lambda_0(t)e^{U'\gamma}} \frac{(\Lambda_0(t)e^{U'\gamma})^k}{k!} \frac{1}{\Gamma(\alpha k)} \int_0^x e^{-y} y^{\alpha k-1}\,dy.$$

This implies that β acts like a scale parameter in the distribution of $\mathcal{S}(t)$.

We are interested in inference concerning the median of $\mathcal{S}(t)$, i.e., the quantity

$$Q(t; \Lambda_0(\cdot), \gamma, \alpha, \beta|U) = \begin{cases} 0 \text{ if } e^{-\Lambda_0(t)e^{U'\gamma}} \geqslant \frac{1}{2} \\ \text{solution of (2.1) below if } e^{-\Lambda_0(t)e^{U'\gamma}} < \frac{1}{2}, \end{cases}$$

where the equation to be solved (for the variable x) is given by

$$\sum_{k=0}^{\infty} e^{-\Lambda_0(t)e^{U'\gamma}} \frac{(\Lambda_0(t)e^{U'\gamma})^k}{k!} \frac{\beta^{\alpha k}}{\Gamma(\alpha k)} \int_0^x e^{-\beta y} y^{\alpha k-1}\, dy = \frac{1}{2}. \tag{2.1}$$

Consequently,

$$Q(t;\Lambda_0(\cdot),\gamma,\alpha,\beta|U) = \beta \bar{Q}(\Lambda_0(t)e^{U'\gamma},\alpha),$$

where

$$\bar{Q}(s,\alpha) = \begin{cases} 0 \text{ if } s \leqslant \ln(2), \\ \text{solution of (2.2) below if } s > \ln(2), \end{cases}$$

where the equation to be solved (for x) is given by

$$\sum_{k=0}^{\infty} e^{-s} \frac{s^k}{k!} \frac{1}{\Gamma(\alpha k)} \int_0^x e^{-y} y^{\alpha k-1}\, dy = \frac{1}{2}. \tag{2.2}$$

Let

$$G(x,y,\alpha) = \sum_{k=0}^{\infty} e^{-y} \frac{y^k}{k!} \frac{1}{\Gamma(\alpha k)} \int_0^x e^{-s} s^{\alpha k-1}\, ds - \frac{1}{2}.$$

Note that $G(x,y,\alpha)$ is strictly increasing and continuous in x and when $y > \ln 2$, for all α, we have

$$\lim_{x\longrightarrow 0} G(x,y,\alpha) = e^{-y} - \frac{1}{2} < 0 \text{ and } \lim_{x\longrightarrow \infty} G(x,y,\alpha) = \frac{1}{2} > 0.$$

Consequently, (2.2) has unique solution in x when $y > \ln 2$.

2.3 *Modeling in presence of terminal events*

Let us fix notation first. As before, let U denote the covariate vector of interest and Z the underlying frailty capturing the dependence across various processes. Let $(U, Z, C, D, N(\cdot), \mathcal{S}(\cdot))$ denote various processes of interest, where D denotes death, C denotes the independent censoring variable, Z a frailty variable. We assume that given (U, Z), the processes $(C, D, N(\cdot))$ are mutually independent. The total observation period is denoted by $[0, \tau]$. Recall that the intensity λ for the recurrent event process is given by

$$\lambda(t) = z\lambda_0(t)e^{U'\gamma},\ \Lambda_0(t) = \int_0^t \lambda_0(s)\, ds \text{ with } \Lambda_0(\tau) = 1.$$

Moreover, we also assume

$$E(Z|N) = E(Z)$$

and the model for the death process is given by

$$h(t) = zh_0(t)e^{U'\eta}.$$

We also denote

$$\widetilde{D} = \min\{C, D, \tau\}, \Delta_i = I(D_i \leqslant \widetilde{D}), \mu_Z = E(Z) \text{ and } \bar{U} = (1, U).$$

We adapt the Huang and Wang (2004) approach for estimation for $(\Lambda_0, H_0, \gamma, \eta)$. We present them below for sake of completeness. They proposed the following estimate $\hat{\Lambda}_0(t)$ for $\Lambda_0(t)$,namely,

$$\hat{\Lambda}_0(t) = \prod_{s_{(l)} > t} \left(1 - \frac{d_{(l)}}{R_{(l)}}\right),$$

where $\{s_{(l)}\}$ are the ordered and distinct values of event times $\{t_{ij}\}$ and

$$d_{(l)} = \#\text{ of events at } \ s_{(l)} \text{ and } R_{(l)} = \sum_{i,j} I\{t_{ij} \leqslant s_{(l)} \leqslant \widetilde{D}_i\}.$$

Adapting Huang and Wang (2004) approach, estimates for $(\log \mu_Z, \gamma) = \tilde{\gamma}$ are obtained through solving

$$\frac{1}{n}\sum_{i=1}^{n} w_i \bar{U}_i' \Big[m_i\{\hat{\Lambda}_0(\widetilde{D}_i)\}^{-1} - e^{\bar{U}_i'\tilde{\gamma}}\Big] = 0,$$

where $m_i = N_i(\tilde{D}_i)$ is the total number of observed events for individual i. We next present the estimates for the death model as proposed in Huang and Wang (2004). Denote

$$\hat{Z}_i = \frac{m_i}{\hat{\Lambda}_0(\widetilde{D}_i)e^{U_i'\hat{\gamma}}}.$$

Estimate of η is given by the solution of the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \Delta_i \left\{U_i - \frac{\sum_{j=1}^{n} U_j \hat{Z}_j e^{U_j'\eta} I(\widetilde{D}_j \geqslant \widetilde{D}_i)}{\sum_{j=1}^{n} \hat{Z}_j e^{U_j'\eta} I(\widetilde{D}_j \geqslant \widetilde{D}_i)}\right\} = 0$$

and

$$\hat{H}_0(t) = \sum_{1 \leqslant i \leqslant n, \widetilde{D}_i \leqslant t} \frac{\Delta_i}{\sum_{k=1}^{n} \hat{Z}_k e^{U_k'\tilde{\eta}} I(\widetilde{D}_k \geqslant \widetilde{D}_i)}.$$

We will now present our estimate for the median cost in the presence of death, where only $S(t \wedge C \wedge D)$ is observable. Note that incurring medical cost beyond death is not possible. So, to identify the median cost in this setup, the appropriate cost process is $\mathcal{S}(t \wedge D)$. So,

$$P(\mathcal{S}(t \wedge D) \leqslant x|U) = P(N(t \wedge D) = 0) + \sum_{k=1}^{\infty} P(N(t \wedge D)) = k|U)F_{\mathcal{S}}(x; \alpha k, \beta),$$

where $F_{\mathcal{S}}(x; \alpha k, \beta) = \beta^{\alpha k}/\Gamma(\alpha k) \int_0^x e^{-\beta y} y^{\alpha k - 1} dy$, is the cumulative distribution function of a Gamma$(\alpha k, \beta)$ distributed random variable. We want the median to be dependent on the covariates and so we take

$$\begin{aligned} P(\mathcal{S}(t \wedge D) \leqslant x|U) &= \int P(\mathcal{S}(t \wedge D) \leqslant x|U, Z) dF_Z(z) \\ &= \int P(\mathcal{S}(t) \leqslant x, t \leqslant D|U, Z) dF_Z(z) + \int P(\mathcal{S}(t) \leqslant x, t > D|U, Z) dF_Z(z) = I + II. \end{aligned}$$

Note now that due to the assumption of independence of $(C, D, N(\cdot))$ given (U, Z), we have

$$\begin{aligned} I &= \int \int_t^{\infty} P(\mathcal{S}(t) \leqslant x|D = d, U, Z) dF_D(d|U, Z) dF_Z(z) \\ &= \int \int_t^{\infty} P(\mathcal{S}(t) \leqslant x|U, Z) dF_D(d|U, Z) dF_Z(z) \\ &= \int P(\mathcal{S}(t) \leqslant x|U, Z) S_D(t|U, Z) dF_Z(z). \end{aligned}$$

Similarly,

$$\begin{aligned} II &= \int \int_0^t P(\mathcal{S}(D) \leqslant x|D = d, U, Z) dF_D(d|U, Z) dF_Z(z) \\ &= \int \int_0^t P(\mathcal{S}(d) \leqslant x|U, Z) dF_D(d|U, Z) dF_Z(z). \end{aligned}$$

So, the median at time t is given by the solution to the equation

$$\int P(\mathcal{S}(t) \leqslant x|U, Z) S_D(t|U, Z) dF_Z(z) + \int \int_0^t P(\mathcal{S}(d) \leqslant x|U, Z) dF_D(d|U, Z) dF_Z(z) = \frac{1}{2}.$$

Focusing on $P(\mathcal{S}(t) \leqslant x|U, Z)$ first, we note

$$\begin{aligned} P(\mathcal{S}(t) \leqslant x|U, Z) &= P(N(t) = 0|U, Z) + \sum_{k=1}^{\infty} P(N(t) = k|U, Z) F_{\mathcal{S}}(x; \alpha k, \beta, U) \\ &= \left\{ e^{-Z\Lambda_0(t)e^{U'\gamma}} + \sum_{k=1}^{\infty} e^{-Z\Lambda_0(t)e^{U'\gamma}} \frac{(Z\Lambda_0(t)e^{U'\gamma})^k}{k!} F_{\mathcal{S}}(x; \alpha k, \beta, U) \right\}. \end{aligned}$$

Taking expected value we have the equation

$$E_Z\left[\{e^{-Z\Lambda_0(t)e^{U'\gamma}}+\sum_{k=1}^{\infty}e^{-Z\Lambda_0(t)e^{U'\gamma}}\frac{(Z\Lambda_0(t)e^{U'\gamma})^k}{k!}F_{\mathcal{S}}(x;\alpha k,\beta,U)\}S_D(t|U,Z)\right]$$
$$+E_Z\left[\int_0^t\{e^{-Z\Lambda_0(d)e^{U'\gamma}}+\sum_{k=1}^{\infty}e^{-Z\Lambda_0(d)e^{U'\gamma}}\frac{(Z\Lambda_0(d)e^{U'\gamma})^k}{k!}F_{\mathcal{S}}(x;\alpha k,\beta,U)\}dF_D(d|U,Z)\right]$$
$$=\frac{1}{2}. \tag{2.3}$$

Finally, we replace all quantities by their respective estimates to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left[\{e^{-\hat{Z}_i\hat{\Lambda}_0(t)e^{U'\hat{\gamma}}}+\sum_{k=1}^{\infty}e^{-\hat{Z}_i\hat{\Lambda}_0(t)e^{U'\hat{\gamma}}}\frac{(\hat{Z}_i\hat{\Lambda}_0(t)e^{U'\hat{\gamma}})^k}{k!}F_{\mathcal{S}}(x;\hat{\alpha}k,\hat{\beta},U)\}\hat{S}_D(t|U,\hat{Z}_i)\right]$$
$$+\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\{e^{-\hat{Z}_i\hat{\Lambda}_0(\widetilde{D}_j)e^{U'\hat{\gamma}}}+\right.$$
$$\left.\sum_{k=1}^{\infty}e^{-\hat{Z}_i\hat{\Lambda}_0(\widetilde{D}_j)e^{U'\hat{\gamma}}}\frac{(\hat{Z}_i\hat{\Lambda}_0(\widetilde{D}_j)e^{U'\hat{\gamma}})^k}{k!}F_{\mathcal{S}}(x;\hat{\alpha}k,\hat{\beta},U)\}I(\widetilde{D}_j\leqslant t)d\hat{F}_D(\widetilde{D}_j|U,\hat{Z}_i))\right]$$
$$=\frac{1}{2}.$$

Consequently, solving the above equation for x yields the estimate for the median $Q(t;\Lambda_0(\cdot),H_0(\cdot),\gamma,\eta,\alpha,\beta|U)$.

Remark: In practice, direct plugin of $\hat{Z}_1,\cdots,\hat{Z}_n$ did not yield good numerical results. However, computing the expectation with respect to the distribution of Z in equation (2.3) by estimating its density f_Z by using Kernel density estimation of $\hat{Z}_1,\cdots,\hat{Z}_n$ yielded much improved results. We applied the methods by Sheather and Jones (1991) for bandwidth selection and Wand and Jones (1994) for kernel density estimation.

3. Asymptotic Results

We first focus on the independent right censoring case discussed in Section 2.2. Let $\hat{\Lambda}_0,\hat{\alpha},\hat{\beta},\hat{\gamma}$ be the estimates of the corresponding parameters based on observed data. We can then estimates

$Q(t;\Lambda_0(\cdot),\alpha,\beta,\gamma|U)$ by

$$\hat{Q}(t) = Q(t;\hat{\Lambda}_0(\cdot),\hat{\alpha},\hat{\beta},\hat{\gamma}|U) = \hat{\beta}\bar{Q}(\hat{\Lambda}_0(t)e^{U'\gamma},\hat{\alpha}).$$

In order to find the asymptotic distribution of $\hat{Q}$, we apply the delta method to the joint asymptotic distribution of $(\hat{\Lambda},\hat{\alpha},\hat{\beta},\hat{\gamma})$. An estimating equation approach for estimating $(\hat{\Lambda}_0(\cdot),\hat{\gamma})$ was proposed in Wang *and others* (2001), where they also obtained asymptotic normality. Following the same proof, one can deduce joint asymptotic normality provided one finds the corresponding covariance term. The estimators $(\hat{\alpha},\hat{\beta})$ are separately obtained from the cost data and they are also asymptotically normal. Thus, $(\hat{\Lambda}_0,\hat{\alpha},\hat{\beta},\hat{\gamma})$ are jointly asymptotically normal with the cross-covariance between $(\hat{\Lambda}_0(\cdot),\hat{\gamma})$ and $(\hat{\alpha},\hat{\beta})$ being zero. Thus we need to calculate the partial derivatives $\frac{\partial Q}{\partial l}, \frac{\partial Q}{\partial \gamma}, \frac{\partial Q}{\partial \alpha}, \frac{\partial Q}{\partial \beta}$ where $l = \Lambda_0(t)$ and evaluate at the true values of $(\Lambda_0(t),\gamma,\alpha,\beta)$ to obtain asymptotic variance of $\hat{Q}$. Since the true values of these parameters are unknown, we will finally substitute the corresponding estimates $\hat{\Lambda}_0,\hat{\gamma},\hat{\alpha},\hat{\beta}$. A calculation of derivatives yield

$$\frac{\partial Q}{\partial l} = \beta\left(\frac{\partial \bar{Q}(y,\alpha)}{\partial y}|_{y=le^{U'\gamma}}\right)e^{U'\gamma}, \frac{\partial Q}{\partial \gamma} = \beta\left(\frac{\partial \bar{Q}(y,\alpha)}{\partial y}|_{y=le^{U'\gamma}}\right)le^{U'\gamma}z$$
$$\frac{\partial Q}{\partial \alpha} = \beta\frac{\partial \bar{Q}(y,\alpha)}{\partial \alpha}|_{y=le^{U'\gamma}} \text{ and } \frac{\partial Q}{\partial \beta} = \bar{Q}(le^{U'\gamma},\alpha).$$

Thus, we would like to compute $\frac{\partial \bar{Q}(y,\alpha)}{\partial y}, \frac{\partial \bar{Q}(y,\alpha)}{\partial \alpha}$. As $\bar{Q}(y,\alpha)$ is defined implicitly as the solution of (2.2) (or equivalently, $G(x,y,\alpha)=0$), we apply the implicit function theorem to calculate the derivatives. This yields

$$\frac{\partial \overline{Q}}{\partial y} = \frac{\frac{\partial G}{\partial y}|_{(\overline{Q}(y,\alpha),y,\alpha)}}{\frac{\partial G}{\partial x}|_{(\overline{Q}(y,\alpha),y,\alpha)}} \text{ and } \frac{\partial \overline{Q}}{\partial \alpha} = \frac{\frac{\partial G}{\partial \alpha}|_{(\overline{Q}(y,\alpha),y,\alpha)}}{\frac{\partial G}{\partial x}|_{(\overline{Q}(y,\alpha),y,\alpha)}}.$$

Now we can easily compute $\frac{\partial G}{\partial y}$ as

$$\frac{\partial G}{\partial y} = \sum_{k=0}^{\infty} e^{-y}\frac{y^k}{k!}\left[\frac{1}{\Gamma(\alpha(k+1))}I(x;\alpha(k+1)) - \frac{1}{\Gamma(\alpha k)}I(x;\alpha k)\right],$$

where $I(x;r) = \int_0^x e^{-u}u^{r-1}\,du$ is the incomplete gamma integral and $\frac{I(x;r)}{\Gamma(r)} = 1$ if $r = 0$. Likewise,

$$\frac{\partial G}{\partial \alpha} = \sum_{k=1}^{\infty} e^{-y}\frac{y^k}{k!}\frac{1}{\Gamma(\alpha k)}\left[k\int_0^x e^{-u}u^{\alpha k-1}\ln u\,du - \frac{\Gamma'(\alpha k)}{\Gamma(\alpha k)}I(x;\alpha k)\right]$$

and

$$\frac{\partial G}{\partial x} = \sum_{k=1}^{\infty} e^{-y} \frac{y^k}{k!} \frac{1}{\Gamma(\alpha k)} e^{-x} x^{\alpha k - 1}.$$

Plugging in, and recalling $l = \Lambda_0(t)$ and denoting $\hat{y} = \hat{\Lambda}_0(t) e^{U'\hat{\gamma}}$, we obtain

$$\begin{pmatrix} \hat{J}_1 \\ \hat{J}_2 \\ \hat{J}_3 \\ \hat{J}_4 \end{pmatrix} := \begin{pmatrix} \frac{\partial Q}{\partial l} \\ \frac{\partial Q}{\partial \gamma} \\ \frac{\partial Q}{\partial \alpha} \\ \frac{\partial Q}{\partial \beta} \end{pmatrix}_{(\hat{\Lambda}_0(t), \hat{\gamma}, \hat{\alpha}, \hat{\beta})} = \begin{pmatrix} \frac{\hat{\beta} e^{U'\hat{\gamma}} \sum_{k=0}^{\infty} \frac{\hat{y}^k}{k!} \left\{ \frac{I(\bar{Q}(\hat{y},\hat{\alpha});\hat{\alpha}(k+1))}{\Gamma(\hat{\alpha}(k+1)} - \frac{I(\bar{Q}(\hat{y},\hat{\alpha});\hat{\alpha}k)}{\Gamma(\hat{\alpha}k} \right\}}{\sum_{k=1}^{\infty} \frac{\hat{y}^k}{k!} \frac{e^{-\bar{Q}(\hat{y},\hat{\alpha})}}{\Gamma(\hat{\alpha}k)} [\bar{Q}(\hat{y},\hat{\alpha})]^{\hat{\alpha}k-1}} \\ \frac{z\hat{\Lambda}_0(t)\hat{\beta} e^{U'\hat{\gamma}} \sum_{k=0}^{\infty} \frac{\hat{y}^k}{k!} \left\{ \frac{I(\bar{Q}(\hat{y},\hat{\alpha});\hat{\alpha}(k+1))}{\Gamma(\hat{\alpha}(k+1)} - \frac{I(\bar{Q}(\hat{y},\hat{\alpha});\hat{\alpha}k)}{\Gamma(\hat{\alpha}k} \right\}}{\sum_{k=1}^{\infty} \frac{\hat{y}^k}{k!} \frac{e^{-\bar{Q}(\hat{y},\hat{\alpha})}}{\Gamma(\hat{\alpha}k)} [\bar{Q}(\hat{y},\hat{\alpha})]^{\hat{\alpha}k-1}} \\ \frac{\hat{\beta} \sum_{k=1}^{\infty} \frac{\hat{y}^k}{k!} \left\{ \frac{\hat{\alpha}\Gamma'(\hat{\alpha}k) I(Q(\hat{y},\hat{\alpha}))}{\Gamma^2(\hat{\alpha}k)} - \frac{\int_0^{\bar{Q}(\hat{y},\hat{\alpha})} e^{-u} u^{\hat{\alpha}k-1} \ln u \, du}{\Gamma(\hat{\alpha}k)} \right\}}{\sum_{k=1}^{\infty} \frac{\hat{y}^k}{k!} \frac{1}{\Gamma(\hat{\alpha}k)} e^{-\bar{Q}(\hat{y},\hat{\alpha})} [\bar{Q}(\hat{y},\hat{\alpha})]^{\hat{\alpha}k-1}} \\ \bar{Q}(\hat{y},\hat{\alpha}) \end{pmatrix} \tag{3.4}$$

Let as in Wang *and others* (2001),

$$b(t) = \sum_{j=1}^{N(t)} \left\{ \frac{\int_t^{T_0} \mathbf{1}(t_j \leqslant u \leqslant Y) dQ(u)}{R^2(u)} - \frac{\mathbf{1}\{t < t_j \leqslant T_0\}}{R(t_j)} \right\}$$

and

$$e = \int \frac{w\bar{x}' m b(y) dV(w, \bar{x}, m, y)}{F(y)} + w\bar{x}'[mF^{-1}(y) - e^{\bar{x}'\gamma}.$$

Let $\Sigma = \mathcal{D}(e)$ be the dispersion matrix and let $\Psi = E\left[-\frac{\partial e}{\partial \gamma}\right]$. Then,

$$\sqrt{n}(\hat{\gamma} - \gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi^{-1} e_i + o_p(1).$$

Consequently, the asymptotic dispersion matrix for $\begin{pmatrix} \hat{\Lambda}_0(t) \\ \hat{\gamma} \end{pmatrix}$ is given by

$$\begin{pmatrix} E(d^2(t)) & E(d(t)e(t))'(\Psi')^{-1} \\ \Psi^{-1}E(d(t)e(t)) & \Psi^{-1}\Sigma(\Psi')^{-1} \end{pmatrix}, \text{ where } d(t) = F(t)[c + \Lambda_0(T_0)b(t).$$

If N i.i.d. samples from Gamma(α, β) distribution are $R_1, \cdots, R_N$ and $(\hat{\alpha}, \hat{\beta})$ is the maximum likelihood estimate of (α, β), then $(\hat{\alpha}, \hat{\beta})$ has asymptotic dispersion matrix

$$\frac{1}{N} \begin{pmatrix} \frac{(\Gamma'(\alpha))^2 - \Gamma''(\alpha)\Gamma(\alpha)}{\Gamma^2(\alpha)} & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}.$$

In our case, for n individual, we have $N = \sum_{i=1}^{n} N_i(T_0)$ and the cost data is

$$\omega_{11}, \cdots, \omega_{1m_1}; \cdots, \omega_{n_1}, \cdots, \omega_{nm_n}.$$

Let $\frac{N}{n} \to \xi$ in probability and let $\hat{\xi} = \frac{N}{n}$. The variance-covariance matrix can then be approximated by

$$\frac{1}{n}\frac{n}{N}\begin{pmatrix} \frac{(\Gamma'(\hat{\alpha}))^2 - \Gamma''(\hat{\alpha})\Gamma(\hat{\alpha})}{\Gamma^2(\hat{\alpha})} & \frac{1}{\hat{\beta}} \\ \frac{1}{\hat{\beta}} & \frac{\hat{\alpha}}{\hat{\beta}} \end{pmatrix}.$$

Thus, the asymptotic dispersion matrix of $(\hat{\Lambda}_0(t), \hat{\gamma}, \hat{\alpha}, \hat{\beta})$ can be approximated by

$$\frac{1}{n}\begin{pmatrix} \hat{E}(d^2(t)) & \hat{E}[d(t)e(t)'(\Psi')^{-1} & 0 & 0 \\ \hat{E}[d(t)\Psi^{-1}e(t) & \hat{\Psi}^{-1}\hat{\Sigma}\hat{\Psi}^{-1} & 0 & 0 \\ 0 & 0 & \frac{n\{(\Gamma'(\hat{\alpha}))^2 - \Gamma''(\hat{\alpha})\Gamma(\hat{\alpha})\}}{N\Gamma^2(\hat{\alpha})} & \frac{n}{N\hat{\beta}} \\ 0 & 0 & \frac{n}{N\hat{\beta}} & \frac{n\hat{\alpha}}{N\hat{\beta}^2} \end{pmatrix},$$

where $\hat{E}(\cdot), \hat{\Psi}, \hat{\Sigma}$ are bootstrap estimates as defined in Wang *and others* (2001). Thus, $\hat{Q}(t)$ is asymptotically normal with mean $Q(t)$ and its asymptotic variance can be approximated by

$$\frac{1}{n}\left\{\hat{J}_1^2\hat{E}(d^2(t)) + 2\hat{J}_1\hat{E}[d(t)e(t)'(\Psi')^{-1}]\hat{J}_2 + \hat{J}_2'\hat{\Psi}^{-1}\hat{\Sigma}\hat{\Psi}^{-1}\right\}$$
$$+ \frac{1}{N}\left\{\hat{J}_3^2\frac{\Gamma'(\hat{\alpha})^2 - \Gamma''(\hat{\alpha})\Gamma(\hat{\alpha})}{\Gamma^2(\hat{\alpha})} + 2\frac{\hat{J}_3\hat{J}_4}{\hat{\beta}} + \frac{\hat{J}_4^2\hat{\alpha}}{\hat{\beta}^2}\right\}.$$

The above can be used to test hypothesis and construct confidence interval about $Q(\cdot)$. All this holds provided $\Lambda_0(t)e^{U'\gamma} > \ln 2$, or equivalently, median is the unique solution of (2.2). This will hold provided $|z|$ is bounded and t is appropriately large, which we will assume.

Remark 1: The asymptotic distribution in presence of terminal events follows similar steps to the independent censoring case. Using the asymptotic properties of the underlying parameters $(\hat{\Lambda}_0, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\eta}, \hat{H}_0)$ established in Huang and Wang (2004) and the delta method, we can use similar steps as before and establish the asymptotic results under reasonable regularity conditions.

Remark 2: In practice, the covariance structure established through asymptotic theory has been shown to perform poorly due to the complexity of the structure, which requires various numerical approximations reducing the efficiency, see Wang *and others* (2001) and Huang and Wang (2004). Consequently, we used bootstrapping to estimate the variance in our inference procedure.

4. Simulation Studies

We conducted simulations to investigate the finite sample properties of the proposed estimates for the median cost process. We generated data from a recurrent event process with the following intensity process

$$\lambda(t) = z\lambda_0(t)\exp(U'\gamma) \quad \text{with} \quad \lambda_0(t) = \frac{1}{4}t^{-(1/2)},$$

with the covariate $U \sim N(0,1)$ and $z \sim \text{Gamma}(\alpha = 2, \beta = 0.2)$ such that the mean is 10. We generated the survival times for the terminal event from a proportional hazards model with baseline hazards as constant and a covariate effect 0.5 and an administrative censoring of $\tau = 4$. The estimation was investigated for 30% and 50% censoring of the survival time. Further, we generated the individual costs $\omega \sim \text{Gamma}(\alpha, \beta)$ with $\alpha = 25, \beta = 0.5$. We generated samples of size $n = 250$ and $n = 500$. For each of the samples, we used $m = 200$ number of bootstraps with 1000 replicates. We are not presenting the estimates for the recurrent events and the survival times model as they have been previously investigated in Huang and Wang (2004). Also, the estimates of the parameters for the cost model is obtained by maximization of the gamma likelihood based on complete data. Hence, their finite sample performance is also well-studied in the literature. We focus on the performance of the estimate for the median cost. In Table 1, we present the quantile estimates for the two sample sizes in absence of death, ie. only censoring by $\tau = 4$.

We do find the biases to be small for all sample sizes investigated. The coverage probability based on Wald type confidence intervals appears closer to the nominal level with increasing sample size. In Table 2, we present the results for simulation in presence of terminal events. We compared the performance based on two differing censoring percentages of death, 30% and 50%, ie, 70% were observed dead during the follow-up and 50% were observed dead for the two censoring setups. Again, the biases for estimating the median costs are small irrespective of sample size, with coverage probability getting closer to the nominal level as the sample size increases. Furthermore, the coverage probabilities for the confidence intervals appear reasonable

for both the 30% and 50% censoring.

5. Application to the SEER-Medicare Data

The Surveillance, Epidemiology and End-Results (SEER) -Medicare linked data are population-based data for studying cancer epidemiology and quality of cancer-related health services. The SEER-Medicare linked data consist of a linkage of two large population-based databases, SEER and Medicare. We considered data of cancer incidence diagnosed between 1991 and 2010. The Medicare data contain information on medical costs between 1986 and 2004. The linked data consist of cancer patients in the SEER data who were enrolled in Medicare during the study period of the Medicare data. Details of each data and linkage are discussed in Warren *and others* (2002). Although the linkage criterion sounds simple, it creates a left truncated and right censored sample because the two data sets have different starting times. In the SEER-Medicare linked data, patients diagnosed with cancer before 1991 form a prevalent cohort, because only those patients who survived through 1991 were included. Patients diagnosed with cancer after January $1, 1991$ form an incident cohort, because those patients were recruited at the onset of disease. Patients survived through 2010 were considered censored. Our sample consisted of $n = 33,417$ patients with an average observation window of 1774 days. These patients had in total $87,533$ hospitalizations with an average number of 2.6 hospitalizations per patient.

We present our analysis of repeated hospitalization and accumulated cost over these hospitalizations. We focused on the association between the stage at diagnosis, as determined by AJCC and age at diagnosis on the repeated hospitalization, as well as the accumulated cost over these repeated hospitalizations. We corrected the time from diagnosis for the length of stay in the hospital for each hospitalization as the patient was not at risk for hospitalization during that time. The cost was modeled using Gamma distribution such that the mean of the distribution

was modeled as

$$\log \mu = \xi_1 Age + \sum_{j=1}^{4} \xi_{(j+1)} I\{\text{Stage} = j\} + \xi_6 LOS,$$

where LOS refers to the length of stay in a hospital, measured in days. The repeated hospitalization was modeled using non-homogenous Poisson process with the following covariates $U = (\text{Age}, I\{\text{Stage} = 2\}, I\{\text{Stage} = 3\}, I\{\text{Stage} = 4\})$ and the time to death being modeled as the proportional hazards model with covariate U. We present our results for the cost model, recurrent hospitalizations model as well as the time to death model analysis in Tables 3 and 4, respectively. We used $1,000$ bootstrap samples in our analysis.

The estimated effects of the covariates in Table 3 indicates that the stage of diagnosis significantly increased the mean cost of hospitalization. Furthermore, the length of stay in hospital has significant positive association with the cost of hospitalization as expected. We also found that age had a small negative impact on cost of hospitalization after adjusting for the stage and length of stay.

In our analysis of the repeated hospitalization, we find each of the covariates age, as well as Stage at diagnosis to have significantly positively association with intensity of repeat hospitalization indicating that older women were at higher risk for more number of hospitalization, as well as that Stages 2 through 4 had significantly more number of hospitalizations than Stage 1. Our model for the time to death indicates that the higher stages were associated with increased risk for death, so also was age of diagnosis significantly associated with increased risk of death.

In Figure 1, we present the histogram of the cost associated with each stage of illness. As can be seen, the cost data has highly right skewed, thus indicating a need for looking at more robust measures of centrality than the mean. We made comparisons of the estimated cumulative median cost with the approach of Lin (2000), which models the cumulative mean cost under proportional means model. We applied their method assuming each recurrent event as a terminal event for the subject.

Next, in Figure 2 we compared the Lin estimates for the mean cumulative cost with the median cumulative cost proposed here for a woman diagnosed at average age of 70 years. We also provide the survival distribution for each stage of cancer. Observe that the mean cumulative cost estimate for each stage is much higher than the cumulative median cost estimate, thus confirming the observation that the costs are sensitive to the high values observed in Figure 1. We further observe that the cumulative median cost for the Stages 3 and 4 are much higher than Stages 1 and 2 earlier on. We observe that the cost for Stage 3 crosses over that for Stage 4 at early time, whereas the cumulative median cost for Stages 1 and 2 crosses over that for Stages 3 and 4 at a later time point. This may be a consequence of the difference in survival times seen in Figure 3(c) with the probability of survival longer is much higher among individuals diagnosed with Stages 1 and 2 as compared to individuals diagnosed with Stages 3 and 4. The estimates of Lin (2000) does not capture this cross over, as an underlying assumption of their method is the proportionality across means with respect to the covariates. Finally, we present in Table 5 the median cumulative cost with 95% confidence interval for each stage evaluated at various percentiles of observed times of hospitalizations. The estimates do indicate overall increasing costs by stage. They further indicate a significantly higher cost than Stage 1 for Stage 4 from 1.4 years and that of Stages 2 and 3 from approximately 4 years.

6. Discussion

In this article, we proposed a regression analysis approach to recurring medical cost with incomplete follow-up data. This conceptually simple regression model is semiparametric in the sense that the underlying intensity model for the recurrent times are completely unspecified. We further focused on more robust quantile process estimation rather than the usual mean process. An advantage is that one can easily use the proposed median process estimation to estimate the higher or lower order quantiles (example, 5th percentile, 95th percentile etc). This allows us to

assess the effects of covariates on the higher or lower quantile values (example, high expenders or low expenders) rather than on the mean process where the covariate effects may be averaged over. Additionally, we have proposed an approach that tracks expenses with random recurrent times rather than at fixed ad hoc scheduled time.

Our proposed method is computationally easy to implement and exhibited good finite sample properties. In our real data analysis, we focused on the incident part of the data collected by SEER-Medicare program. However, it would be interesting to also include the prevalent part of the data, which leads to the terminal time being left truncated and right censored, causing the cost process to be length biased and informatively censored. Finally, we have used a parametric model to estimate the costs. We did find gamma distribution to be a reasonable choice but it would be interesting to fit a semi-continuous distribution for the costs. Finally, we focused on the cost associated with hospitalizations. However, these methods can be easily adapted to handle other processes like tumor sizes associated with recurrent events.

Funding

This work was supported by the Intramural Research Program of the U.S., National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

Acknowledgment

The author acknowledges that this study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Maryland (`http://biowulf.nih.gov`). *Conflict of Interest*: None declared.

References

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

Bang, H. and Tsiatis, A. A. (2002). Median regression with censored cost data. *Biometrics* **58**(3), 643–649.

Ghosh, D. and Lin, D. Y. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics* **59**(4), 877–885.

Huang, C.-Y. and Wang, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association* **99**(468), 1153–1165.

Huang, Y. (2002). Calibration regression of censored lifetime medical cost. *Journal of the American Statistical Association* **97**(457), 318–327.

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

Lin, D. Y. (2000*b*). Proportional means regression for censored medical costs. *Biometrics* **56**(3), 775–778.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(4), 711–730.

Lin, D. Y., Wei, L. J. and Ying, Z. (2001). Semiparametric transformation models for point processes. *Journal of the American Statistical Association* **96**(454), 620–628.

Luo, X., Huang, C.-Y. and Wang, L. (2013). Quantile regression for recurrent gap time data. *Biometrics* **69**(2), 375–385.

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**(423), 811–820.

Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**(2), 373–379.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 683–690.

Strawderman, R. L. (2000). Estimating the mean of an increasing stochastic process at a censored stopping time. *Journal of the American Statistical Association* **95**(452), 1192–1208.

Sundaram, R. (2007). Robust estimation for analyzing recurrent-event data in the presence of terminal events. *Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*, 245–264.

Wand, M. P and Jones, M. C. (1994). *Kernel Smoothing*. London, UK: Chapman and Hall/CRC.

Wang, M.-C., Qin, J. and Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association* **96**(455), 1057–1065.

Warren, J. L., Klabunde, C. N., Schrag, D., Bach, P. B. and Riley, G. F. (2002). Overview of the seer-medicare data: content, research applications, and generalizability to the united states elderly population. *Medical Care* **40**(8), IV–3.

Table 1. Estimated median cost at 5-, 10-, 25-, 50-, 75-, 90- and 95-th percentiles of event times in absence of terminal event.

		$n = 250$				$n = 500$			
t	$Q(t)$	Bias	SSD	BSD	CP	Bias	SSD	BSD	CP
0.01	0	0	0	0.30	1	0	0	0.01	1
0.04	45.84	−0.06	1.54	1.60	0.971	−0.09	1.09	1.11	0.954
0.25	112.83	0.50	5.16	5.15	0.943	0.23	3.69	3.66	0.949
1.00	241.35	−0.19	7.23	7.05	0.945	0.04	4.95	5.00	0.953
2.25	365.89	−0.16	8.93	8.97	0.948	0.06	6.22	6.37	0.958
3.24	440.98	−0.28	10.01	10.01	0.954	0.04	6.91	7.09	0.959
3.61	465.95	−0.18	10.29	10.34	0.952	0.08	7.14	7.32	0.954

Table 2. Estimated median cost at 5-, 10-, 25-, 50-, 75-, 90- and 95-th percentiles of event times with 30%, 50% and 70% censoring of death.

30% censoring of death									
		$n = 250$				$n = 500$			
t	$Q(t)$	Bias	SSD	BSD	CP	Bias	SSD	BSD	CP
0.002	0	0	0	0	1	0	0	0	1
0.010	0	0	0	0.00	1	0	0	0	1
0.064	47.17	−0.06	2.08	2.18	0.983	−0.32	1.46	1.46	0.957
0.300	99.19	−0.22	5.36	5.85	0.997	−0.98	3.66	3.76	0.979
0.930	156.35	−0.91	9.69	9.84	0.969	−2.37	6.71	6.63	0.942
1.960	195.90	−3.12	12.34	12.05	0.938	−4.66	8.69	8.48	0.929
2.680	208.19	−2.88	12.69	12.49	0.950	−4.70	8.80	8.65	0.925
50% censoring of death									
		$n = 250$				$n = 500$			
t	$Q(t)$	Bias	SSD	BSD	CP	Bias	SSD	BSD	CP
0.005	0	0	0	0	1	0	0	0	1
0.019	0	0.43	3.75	4.78	0.996	0.03	1.00	1.48	1
0.120	57.50	1.10	4.43	4.78	0.960	0.24	2.57	2.88	0.959
0.533	138.69	0.58	9.13	9.35	0.952	−0.58	6.46	6.69	0.957
1.478	214.01	1.83	13.33	13.24	0.949	−0.44	9.06	9.34	0.955
2.637	264.11	0.89	16.01	15.79	0.949	−1.60	10.89	11.16	0.955
3.228	281.48	0.28	17.08	16.64	0.943	−2.29	11.65	11.83	0.942

Table 3. Analysis of the SEER-Medicare Data: Cost Associated with Ovarian Cancer.

Par	Est	SD	95% C.I.
shape: α	1.186	0.009	(1.169, 1.204)
age: ξ_1	-0.008	0.001	(-0.009, -0.007)
stage 1: ξ_2	10.286	0.046	(10.195, 10.377)
stage 2: ξ_3	10.315	0.049	(10.220, 10.410)
stage 3: ξ_4	10.367	0.047	(10.275, 10.458)
stage 4: ξ_5	10.247	0.048	(10.153, 10.340)
LOS: ξ_6	1.377	0.024	(1.330, 1.425)

Table 4. Estimation results for medical costs.

(a) Event process

Par	Est	SD	95% C.I.
age: γ_1	0.032	0.001	(0.030, 0.034)
stage 2: γ_2	0.401	0.046	(0.310, 0.491)
stage 3: γ_3	0.901	0.034	(0.835, 0.968)
stage 4: γ_4	1.172	0.034	(1.105, 1.239)

(b) Death process

Par	Est	SD	95% C.I.
age: η_1	0.054	0.001	(0.052, 0.057)
stage 2: η_2	0.745	0.060	(0.628, 0.861)
stage 3: η_3	1.546	0.044	(1.461, 1.632)
stage 4: η_4	2.083	0.045	(1.994, 2.172)

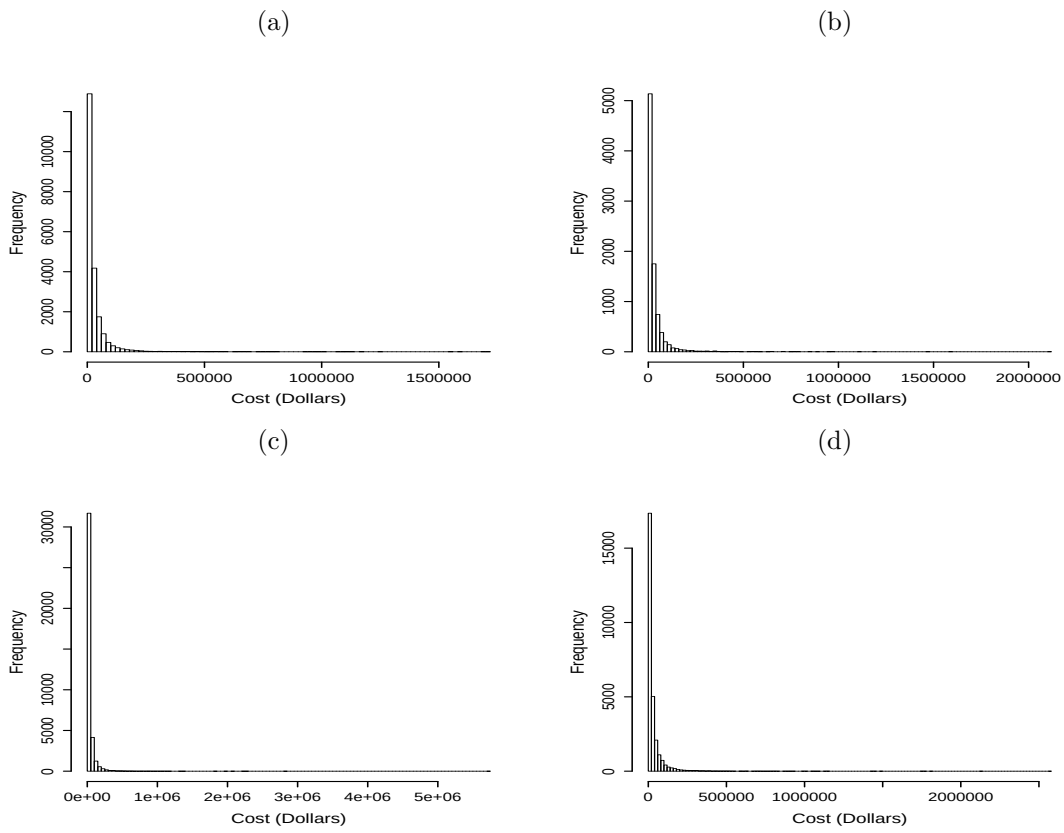

Fig. 1. Histogram of Costs by Stage: (a): Stage 1; (b) Stage 2; (c) Stage 3; (d): Stage 4

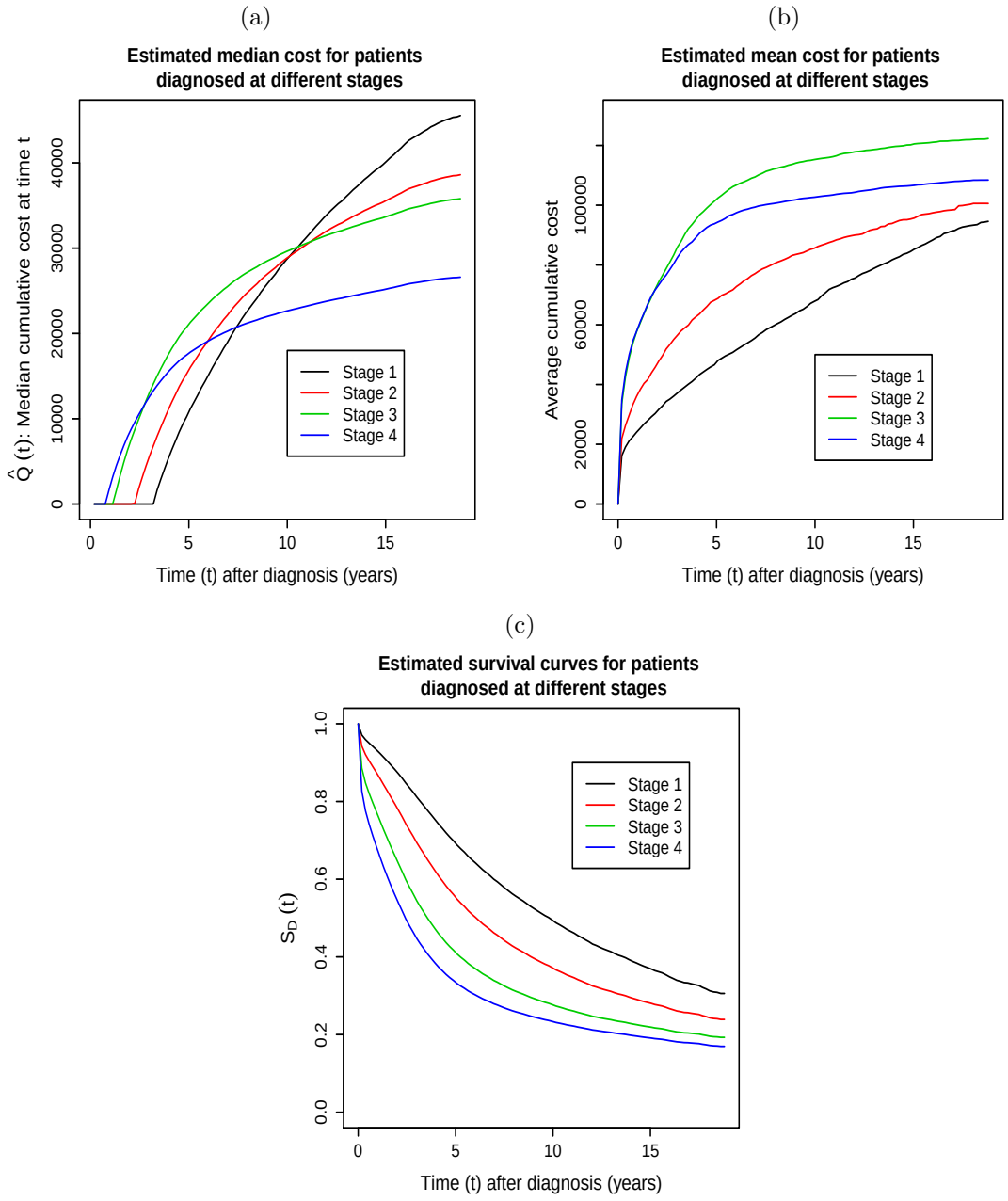

Fig. 2. Comparison of the mean cost process (Lin, 2000) and the median cost process (ours) by stage

Table 5. Estimated median cost at 5-, 10-, 25-, 50-, 75-, 90- and 95-th percentiles of event times for patients diagnosed at age 70 at stages 1–4.

t	$\hat{Q}(t)$	BSD	Wald C.I.
		Diagnosed at stage 1	
0.01	0	0	(0, 0)
0.03	0	0	(0, 0)
0.20	0	0	(0, 0)
1.38	0	0	(0, 0)
3.98	5467.89	1140.37	(3232.76, 7703.21)
7.71	21783.21	1435.39	(18969.85, 24596.58)
10.44	29883.30	1651.10	(26647.15, 33119.45)
		Diagnosed at stage 2	
0.01	0	0	(0, 0)
0.03	0	0	(0, 0)
0.20	0	0	(0, 0)
1.38	0	0	(0, 0)
3.98	11185.42	1305.14	(8627.36, 13743.48)
7.71	24189.58	1612.19	(21029.69, 27349.48)
10.44	29577.06	1781.50	(26085.32, 33068.81)
		Diagnosed at stage 3	
0.01	0	0	(0, 0)
0.03	0	0	(0, 0)
0.20	0	0	(0, 0)
1.38	2216.994	1131.396	(-0.543, 4434.053)
3.98	17631.72	1264.64	(15153.02, 20110.42)
7.71	26798.57	1461.89	(23933.27, 29663.86)
10.44	30063.22	1555.36	(27014.71, 33111.73)
		Diagnosed at stage 4	
0.01	0	0	(0, 0)
0.03	0	0	(0, 0)
0.20	0	0	(0, 0)
1.38	4933.15	973.54	(3025.02, 6841.29)
3.98	15565.83	1116.94	(13376.63, 17755.03)
7.71	21001.93	1239.37	(18572.76, 23431.10)
10.44	22898.22	1303.43	(20343.50, 25452.95)