Bayesian Semiparametric Cure Rate Model with an Unknown Threshold

Guosheng Yin
Assistant Professor
Department of Biostatistics
U. T. M. D. Anderson Cancer Center
Houston, TX

Joint work with LUIS E. NIETO-BARAJAS at ITAM

Outline

- Background and motivation
- Nonparametric prior specification
- Semiparametric model
- Posterior distributions
- Simulation studies
- Example
- Concluding remarks

Cure Rate Model

- In time-to-event analysis for certain diseases, there exists a fraction of the population that is cured of or insusceptible to the disease, who thus will never experience the failure.
- The standard (mixture) cure rate model of Berkson and Gage (1952)

$$S_{pop}(t) = \pi + (1 - \pi)S(t),$$

where $\pi \in (0, 1)$ and $S(t)$ is a proper survival function.

Hazard Functions

- The cumulative hazard function is

$$H_{pop}(t) = -\log S_{pop}(t) = -\log\{\pi + (1-\pi)S(t)\},$$

which satisfies that $\lim_{t\to 0} H_{pop}(t) = 0$ and $\lim_{t\to\infty} H_{pop}(t) = -\log\pi > 0$.

- The hazard rate is

$$h_{pop}(t) = \frac{\mathrm{d}}{\mathrm{d}t} H_{pop}(t) = \frac{(1-\pi)f(t)}{\pi + (1-\pi)S(t)},$$

where $f(t)$ is the density function corresponding to $S(t)$.

Alternative Cure Model

- The alternative cure rate model of Yakolev and Tsodikov (1996) is defined by

$$S_{pop}(t) = \exp\{-\theta F(t)\},$$

where $\theta > 0$ and $F(t)$ is a proper cumulative distribution function.

- This model satisfies the conditions that $\lim_{t\to 0} S_{pop}(t) = 1$ and $\lim_{t\to\infty} S_{pop}(t) = e^{-\theta}$, therefore $S_{pop}(t)$ is not a proper survival function.

Hazard Functions

- The cumulative hazard function is given by

$$H_{pop}(t) = -\log S_{pop}(t) = \theta F(t).$$

- The hazard rate is

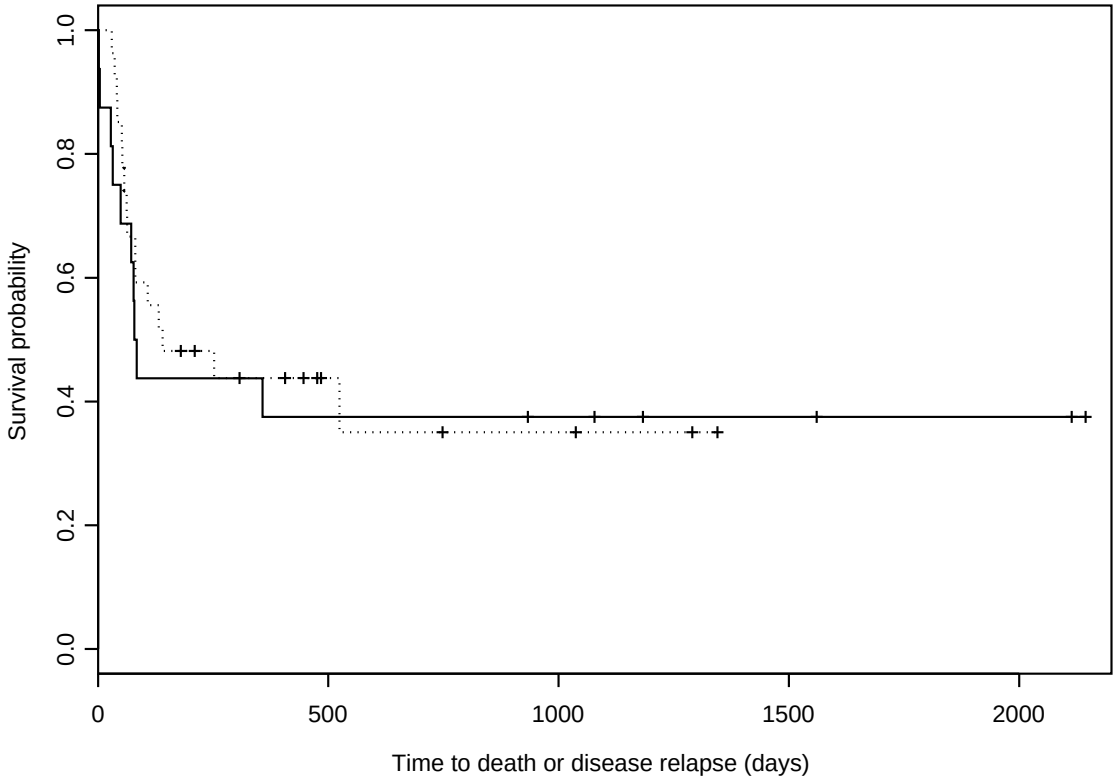
$$h_{pop}(t) = \frac{\mathrm{d}}{\mathrm{d}t} H_{pop}(t) = \theta f(t).$$

- Researchers often model a cure probability, but do not explicitly quantify the finite cure time, which would provide useful information with direct application to the clinical management of the disease.

BMT Example

- Patients had been diagnosed with Hodgkin's disease or with non-Hodgkin's lymphoma.
- Allogeneic BMT (infusion of bone marrow from a matched sibling donor), or autogeneic BMT (reinfusion of the patient's own marrow that had been removed prior to marrow-destroying treatment).
- Of 43 BMT patients, 16 had undergone an allogeneic transplant and 27 an autogeneic transplant.
- To evaluate the leukemia-free survival difference between the two BMT groups.

Figure 1: Kaplan-Meier survival curves stratified by transplant groups: (—) allogeneic group, and (· · ·) autogeneic group.



Conditions for Cure Rate Model

- Conditions required for the hazard function of the entire population in cure rate models:
  (i) $\lim_{t\to 0} H_{pop}(t) = 0$;
  (ii) $\lim_{t\to\infty} H_{pop}(t) = const. < \infty$.
- If we define $h_{pop}(t) = \mathrm{d}H_{pop}(t)/\mathrm{d}t$, a necessary condition for (i) and (ii) to be satisfied is that $\lim_{t\to\infty} h_{pop}(t) = 0$.

An Unknown Threshold

- We propose a new cure rate model for the population hazard rate function that vanishes when t exceeds a certain threshold, say τ,

$$h_{pop}(t) = h(t)I(t \leq \tau),$$

  with $h(t)$ a nonnegative function.

- It is reasonable to let the hazard rate drop to zero after the cure threshold as the subjects who have survived up to τ would become risk-free of the event.

Nonparametric Prior

- We define a nonparametric prior for $h(t)$,

$$h(t) = \sum_{k=1}^{\infty} \lambda_k I(\tau_{k-1} < t \leq \tau_k)$$

- Note that $0 = \tau_0 < \tau_1 < \cdots$ forms a partition of the time axis, and $\lambda_k \sim \text{Ga}(\alpha_k, \beta_k)$.

- Denoting τ_z as the discretized cure time, then $t \leq \tau_z$ can be replaced by $k \leq z$,

$$h_{pop}(t) = \sum_{k=1}^{\infty} \lambda_k I(k \leq z) I(\tau_{k-1} < t \leq \tau_k).$$

Mixture Prior

- If we denote the prior on z by $f(z)$, then the new process $\{\lambda_k^*\}$ with $\lambda_k^* = \lambda_k I(k \leq z)$, can be characterized by

$$f(\lambda_k^*|z) = \text{Ga}(\lambda_k^*|\alpha_k, \beta_k)I(k \leq z) + I(\lambda_k^* = 0)I(k > z).$$

- Marginalizing over z, the prior distribution of λ_k^* is

$$f(\lambda_k^*) = \eta_k \text{Ga}(\lambda_k^*|\alpha_k, \beta_k) + (1 - \eta_k)I(\lambda_k^* = 0),$$

with $\eta_k = \text{P}(z \geq k)$, i.e., λ_k^* has a prior distribution given by a mixture of a gamma distribution and a point mass at zero.

Enhancing Correlations

- To enhance the dependence in the process $\{\lambda_k\}$ we can consider the Markov gamma process of Nieto-Barajas and Walker (2002) for the λ_k's.
- The Markov gamma process $\{\lambda_k\}$ is defined through a latent process $\{u_k\}$:

$$\begin{aligned} \lambda_1 &\sim \text{Ga}(\alpha_1, \beta_1) \\ u_k|\lambda_k &\sim \text{Poi}(c_k\lambda_k) \\ \lambda_{k+1}|u_k &\sim \text{Ga}(\alpha_{k+1} + u_k, \beta_{k+1} + c_k) \end{aligned}$$

for $k = 1, 2, \ldots$.

Cure Fraction

- The cure fraction π is defined as the proportion of the population that will never experience the failure,

$$\pi = \lim_{t \to \infty} S_{pop}(t) = \exp\left\{ -\sum_{k=1}^{z} \lambda_k (\tau_k - \tau_{k-1}) \right\},$$

since λ_k is taken to be zero for $k > z$.

Semiparametric Model

- We propose the hazard function to be of the form

$$h_i(t|\mathbf{x}_i, z_i) = h(t|z_i)e^{\boldsymbol{\gamma}'\mathbf{x}_i(t)},$$

- We assign $h(t|z_i)$ a nonparametric mixture prior to model a cure fraction and a cure time,

$$h(t|z_i) = \sum_{k=1}^{\infty} \lambda_k I(k \leq z_i) I(\tau_{k-1} < t \leq \tau_k),$$

where $\{\lambda_k\}$ is a Markov gamma process common to all subjects and z_i is the cure threshold index for subject i.

Prior on z

- The prior distribution for z, with a support on $\{1, 2, \ldots\}$, can be a positive Poisson distribution, i.e., $z \sim \text{Poi}^+(\mu)$, for $\mu > 0$, that is, if $z \sim \text{Poi}^+(\mu)$ then $z - 1 \sim \text{Poi}(\mu)$.
- To estimate the subject-specific threshold, we assign a prior distribution $f(z_i)$ dependent on the covariates,

$$z_i \sim \text{Poi}^+(e^{\boldsymbol{\delta}' \mathbf{y}_i}),$$

where $\boldsymbol{\delta}$ is a vector of unknown coefficients and $\mathbf{y}_i$ is a $(q+1)$-vector of fixed covariates whose first component is 1.

Cumulative Hazard

- The cumulative hazard function becomes

$$H_i(t|\mathbf{x}_i, z_i) = \sum_{k=1}^{\infty} \lambda_k I(k \le z_i) w_{ki}(t, \mathbf{x}_i, \boldsymbol{\gamma}),$$

where

$$w_{ki}(t, \mathbf{x}_i, \boldsymbol{\gamma}) = \begin{cases} \int_{\tau_{k-1}}^{\tau_k} \exp\{\boldsymbol{\gamma}' \mathbf{x}_i(s)\} \mathrm{d}s, & t > \tau_k \\ \int_{\tau_{k-1}}^{t} \exp\{\boldsymbol{\gamma}' \mathbf{x}_i(s)\} \mathrm{d}s, & t \in (\tau_{k-1}, \tau_k] \\ 0, & \text{otherwise} \end{cases}.$$

- Let $(T_1, \ldots, T_{n_u})$ be the event times, and $(T_{n_u+1}, \ldots, T_n)$ be right-censored.

Likelihood

- The likelihood function is

$$\text{lik}(\boldsymbol{\lambda}, \mathbf{u}, \mathbf{z}, \boldsymbol{\gamma}|\text{data}) = \left\{\prod_{i=1}^{n_u} h_i(t_i|\mathbf{x}_i, z_i)\right\}\left\{\prod_{i=1}^{n} e^{-H_i(t_i|\mathbf{x}_i, z_i)}\right\}$$

$$= \exp\left\{\sum_{i=1}^{n_u} \boldsymbol{\gamma}'\mathbf{x}_i(t_i)\right\} \prod_{k=1}^{\infty}\left[\lambda_k^{r_k} e^{-m_k(\boldsymbol{\gamma},\mathbf{Z})\lambda_k} \prod_{i=1}^{n_u} I(k \le z_i)^{I(\tau_{k-1}<t_i\le\tau_k)}\right],$$

where

$$r_k = \sum_{i=1}^{n_u} I(\tau_{k-1} < t_i \le \tau_k),\ m_k(\boldsymbol{\gamma}, \mathbf{z}) = \sum_{i=1}^{n} I(k \le z_i) w_{ki}(t_i, \mathbf{x}_i, \boldsymbol{\gamma}).$$

Full Conditionals

- The full conditional distributions of λ_k

$$f(\lambda_k|\text{rest}) = \text{Ga}(\lambda_k|\alpha_k+u_{k-1}+u_k+r_k, \beta_k+c_{k-1}+c_k+m_k(\boldsymbol{\gamma}, \mathbf{z})),$$

with $c_0 = 0$ and $u_0 = 0$ w.p.1.

- For u_k, $f(u_k|\text{rest}) \propto$

$$\frac{1}{\Gamma(1+u_k)\Gamma(\alpha_{k+1}+u_k)}\{c_k(\beta_{k+1}+c_k)\lambda_k\lambda_{k+1}\}^{u_k} I_{\{0,1,\ldots\}}(u_k)$$

- For z_i, $f(z_i|\text{rest}) \propto$

$$\exp\left\{-\sum_{k=1}^{\infty}\lambda_k I(k \le z_i) w_{ki}(t_i, \mathbf{x}_i, \boldsymbol{\gamma})\right\} \text{Poi}^+(z_i|e^{\boldsymbol{\delta}'\mathbf{y}_i}) I(k_i \le z_i).$$

Full Conditionals

- The full conditional distribution of $\boldsymbol{\gamma}$

$$f(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathbf{z}, \text{data}) \propto f(\boldsymbol{\gamma}) \exp\left\{\sum_{i=1}^{n_u} \boldsymbol{\gamma}'\mathbf{x}_i(t_i) - \sum_{k=1}^{\infty} m_k(\boldsymbol{\gamma}, \mathbf{z})\lambda_k\right\}.$$

- That of $\boldsymbol{\delta}$ only depends on z_i,

$$f(\boldsymbol{\delta}|\mathbf{z}) \propto f(\boldsymbol{\delta}) \exp\left\{\sum_{i=1}^{n} \left(\boldsymbol{\delta}'\mathbf{y}_i(z_i - 1) - e^{\boldsymbol{\delta}'\mathbf{y}_i}\right)\right\}.$$

- The probability of cure is $\pi_i = \lim_{t\to\infty} S_i(t|\mathbf{x}_i, z_i)$,

$$\pi_i = \exp\left\{-\sum_{k=1}^{z_i} \lambda_k \int_{\tau_{k-1}}^{\tau_k} e^{\boldsymbol{\gamma}'\mathbf{x}_i(s)} \mathrm{d}s\right\}.$$

Simulation I

- We simulated data from the cure rate model

$$S_{pop}(t) = \exp\{-\theta F(t)\}.$$

- We took a triangular distribution Tri$(0, 1, 4)$ as the baseline density, which puts a probability of one to the interval $[0, 4]$ and the mode at 1.
- The censoring time was independently generated from a uniform distribution to yield a 30% censoring rate.
- We took the sample size $n = 100$ and the cure proportion $e^{-\theta} = 0.20$.

Simulation Setup

- We took a fixed time partition with $\tau_0 = 0$ and $\tau_k = \tau_{k-1} + \Delta$, with $\Delta = 0.10$, for $k = 1, \ldots, 100$.
- We considered different values of (α_k, β_k, c_k) and for the hyperprior distribution on μ we took $\mu \sim \text{Ga}(.01, .01)$ to be vague.
- In all scenarios, we ran the Gibbs sampler for 10,000 iterations with a burn-in period of 1,000.
- The logarithm of the pseudo-marginal likelihood (LPML) statistic is used as a model selection criterion, the summary of CPO statistics.

Table 1: Posterior estimates of LPML, cure threshold, posterior mean and 95% credible interval (CI) for π, for a simulated sample of size $n = 100$ with a triangular baseline density.

α_k	β_k	c_k	LPML	$\tau_{\hat{z}}$	$\hat{\pi}$	95% CI
0.01	0.01	0	-153.11	7.1	0.188	(0.110,0.272)
0.1	0.1	0	-149.91	4.2	0.176	(0.105,0.259)
1	1	0	-134.42	3.6	0.133	(0.082,0.193)
1	1	5	-130.35	3.8	0.169	(0.103,0.246)
1	1	20	-127.33	4.0	0.173	(0.098,0.254)
1	1	50	-125.48	4.1	0.178	(0.111,0.259)

Figure 2: Posterior hazard rate estimates (solid line) with 95% credibility intervals (dotted line), $c_k = 0, 5, 20$ and 50.

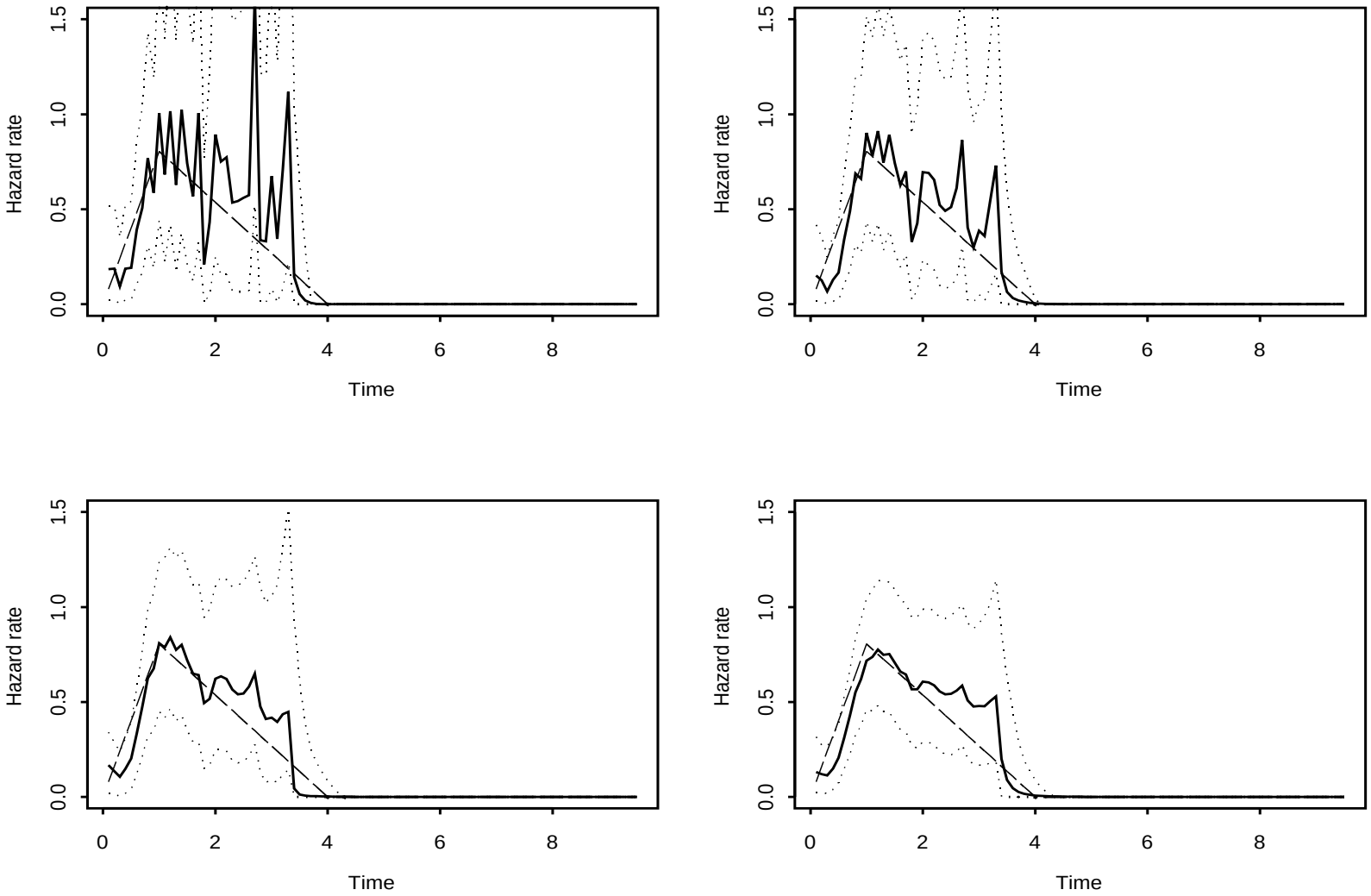


Figure 3: Posterior distributions: μ (left panel) and z (right panel).

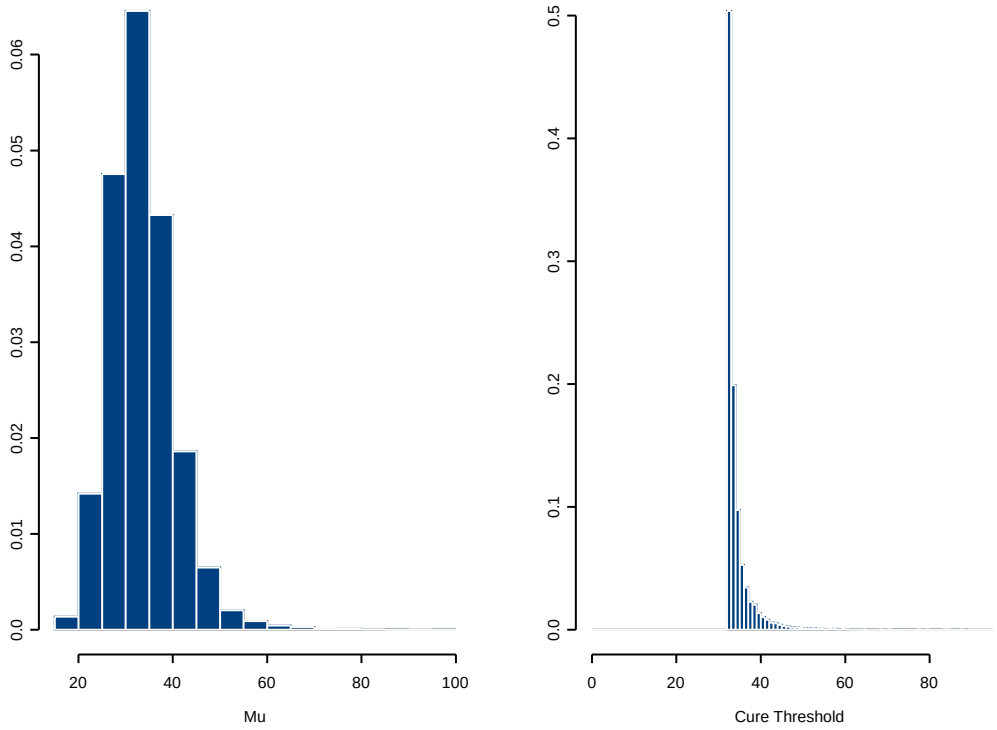
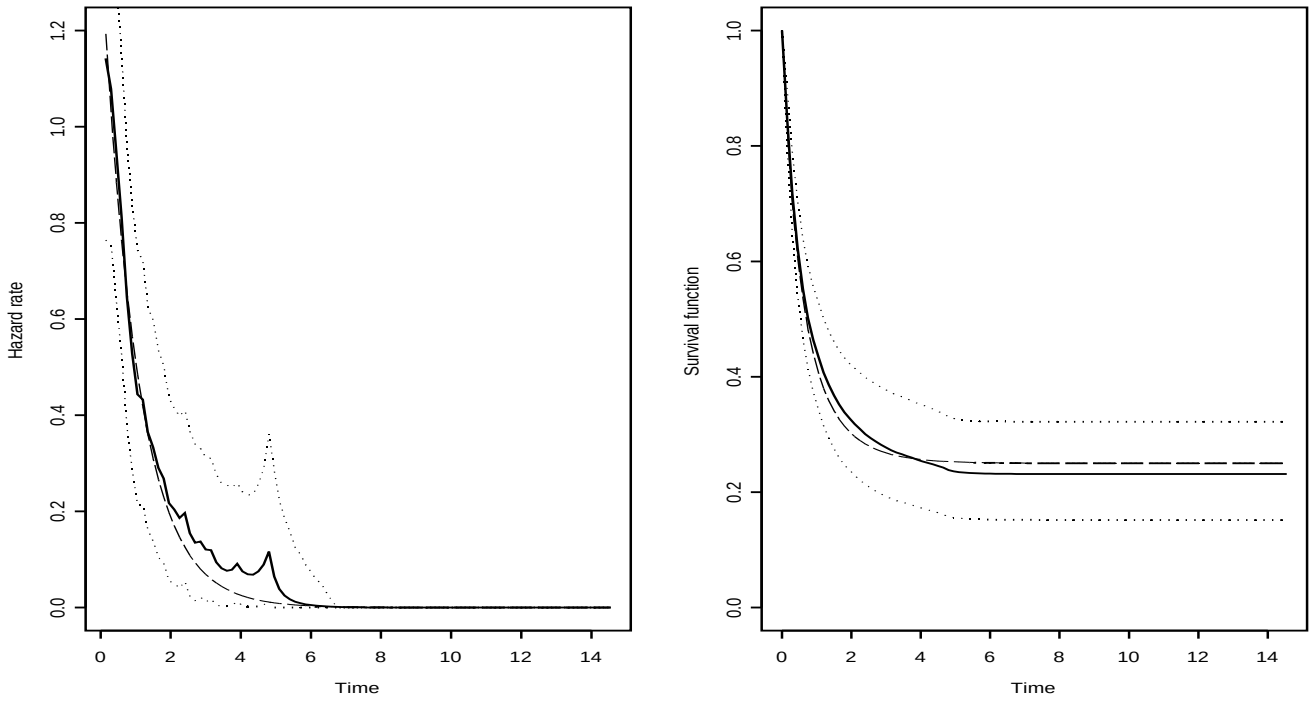


Figure 4: Posterior estimates (solid line) and 95% CI (dotted line): Hazard rate (left panel) and survival function (right panel). The dashed linea are the true functions.



Simulation II

- We simulated data from the same cure model with the baseline density given by an exponential distribution with mean 1.
- The censoring time was independently generated from a uniform distribution to yield 25% of the censoring times.
- We took a sample size of $n = 100$ and a cure proportion $e^{-\theta} = 0.25$.
- We partitioned the time axis by setting $\tau_0 = 0$ and $\tau_k = \tau_{k-1} + \Delta$, with $\Delta = 0.15$, for $k = 1, \ldots, 100$.

Table 2: Posterior estimates of LPML, cure threshold, posterior mean and 95% CI for π, for a simulated sample of size $n = 100$ with an exponential baseline density.

α_k	β_k	c_k	LPML	$\tau_{\hat{z}}$	$\hat{\pi}$	95% CI
0.01	0.01	0	-120.73	10.2	0.249	(0.166,0.340)
0.1	0.1	0	-120.13	5.9	0.229	(0.151,0.317)
1	1	0	-117.40	5.1	0.124	(0.075,0.182)
1	1	5	-111.09	5.4	0.193	(0.127,0.269)
1	1	20	-108.77	5.9	0.219	(0.143,0.302)
1	1	50	-107.47	6.3	0.232	(0.152,0.321)

BMT Data

- The covariates appear in our semiparametric model via two different ways: 1) in a multiplicative manner affecting the "baseline" hazard, and 2) in the Poisson mean of z_i affecting the cure threshold.
- We have also included the estimated covariate effects when fitting the model of Chen, Ibrahim and Sinha. (1999).

Table 3: Posterior estimates of the LPML statistics for the BMT data set.

α_k	β_k	c_k	LPML
1	1	0	-160.25
2	2	0	-159.52
2	2	5	-159.22
2	2	20	-158.21
2	2	50	-158.07
Model of Chen et al.			-165.50

Table 4: Estimated covariate effects in the hazard for the BMT data.

	Our model		Model of Chen, et al.	
Covariate	Mean	95% CI	Mean	95% CI
Intercept	-	-	4.59	(2.79, 6.42)
Trans. type	0.13	(-0.78,1.02)	0.27	(-0.62, 1.14)
Hodgkin	1.20	(0.28,2.22)	1.02	(0.0003, 2.05)
Karnofsky	-0.06	(-0.08,-0.04)	-0.06	(-0.08, -0.04)
Waiting	-0.01	(-0.03,0.004)	-0.01	(-0.03, 0.006)

Table 5: Estimated covariate effects in the cure threshold for the BMT data.

Covariate	Post. Mean	95% CI
Intercept	2.90	(1.19,4.20)
Transplant type	-0.56	(-1.30,0.29)
Hodgkin's disease	-1.10	(-2.08,-0.22)
Karnofsky score	-0.34	(-1.94,1.75)
Waiting time	0.63	(-0.71,1.85)

Table 6: Predictive cure thresholds and cure proportions for new autogeneic transplant patients with Hodgkin's disease ($x_2 = 1$) or non-Hodgkin's lymphoma ($x_2 = 0$) in the BMT data.

Patient	τ_z	π
(x_1, x_2, x_3, x_4)	95% quantile	Post. Mean
$(0, 0, 90, 36)$	14 months	0.40
$(0, 1, 90, 36)$	7 months	0.44
$(0, 0, 60, 36)$	12 months	0.13
$(0, 1, 60, 36)$	6 months	0.14

Figure 5: Predictive hazard estimates for patients with covariates $(0, 0, 90, 36)$ solid line, $(0, 1, 90, 36)$ dotted line, $(0, 0, 60, 36)$ dashed-dotted line and $(0, 1, 60, 36)$ dashed line.

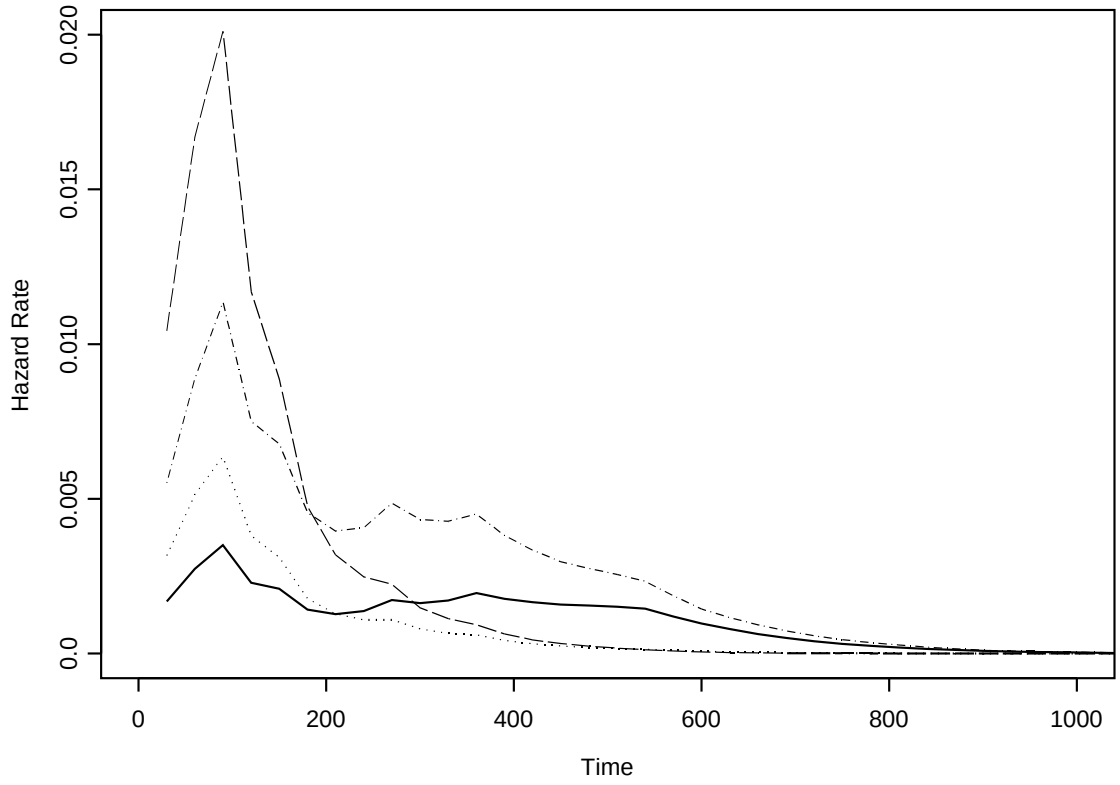
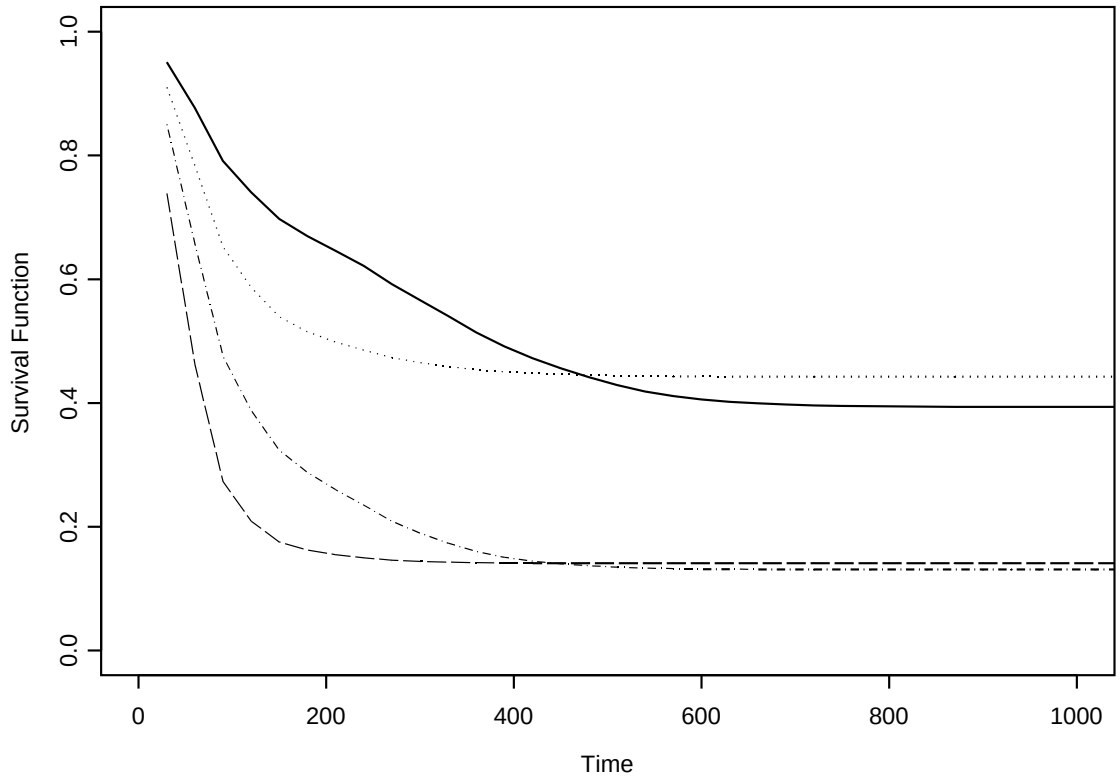


Figure 6: Predictive survival estimates for patients with covariates $(0, 0, 90, 36)$ solid line, $(0, 1, 90, 36)$ dotted line, $(0, 0, 60, 36)$ dashed-dotted line and $(0, 1, 60, 36)$ dashed line.



Summary

- We have proposed a new cure rate model that explicitly incorporates a cure threshold.
- After the cure threshold, the hazard drops to zero, while other cure rate models in the literature allow the hazard to slowly decay to zero.
- We have proposed a mixture prior of a Markov gamma process and a point mass at zero.
- Our semiparametric model uses an exponential link function and allows each patient to have a different cure time depending on covariates.

Thank You!