Statistics
in Medicine

Research Article

Received 7 December 2012, Accepted 31 July 2013 Published online 6 September 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5964

Bayesian random threshold estimation in a Cox proportional hazards cure model

Lili Zhao,[a*†] Dai Feng,[b] Emily L. Bellile[c] and Jeremy M. G. Taylor[a]

In this paper, we develop a Bayesian approach to estimate a Cox proportional hazards model that allows a threshold in the regression coefficient, when some fraction of subjects are not susceptible to the event of interest. A data augmentation scheme with latent binary cure indicators is adopted to simplify the Markov chain Monte Carlo implementation. Given the binary cure indicators, the Cox cure model reduces to a standard Cox model and a logistic regression model. Furthermore, the threshold detection problem reverts to a threshold problem in a regular Cox model. The baseline cumulative hazard for the Cox model is formulated non-parametrically using counting processes with a gamma process prior. Simulation studies demonstrate that the method provides accurate point and interval estimates. Application to a data set of oropharynx cancer patients suggests a significant threshold in age at diagnosis such that the effect of gender on disease-specific survival changes after the threshold. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: threshold; Cox model; cure model; mixture model; Markov chain Monte Carlo

1. Introduction

This paper is motivated by a data set of patients with oropharynx cancer. The clinicians suspect that gender is a prognostic factor for disease-specific survival. However, the Kaplan–Meier (K-M) curves of women and men overlap at early times and cross during the follow-up, which suggests no difference in survival between men and women. But for a cohort of young patients, men tend to have worse survival than women. This seems to indicate a potential threshold in the age of diagnosis such that the gender effect is different below this threshold than above it. Further inspection of the data reveals that the K-M survival curve levels off to about 0.7. The stable plateau at the tail may be taken as empirical evidence of a cured fraction. The use of standard survival analysis for detection of the threshold may be inappropriate as not all patients will die of oropharynx cancer. To this end, we propose a survival model that allows for a threshold in the age of diagnosis to investigate a potential interaction between age and gender when a fraction of patients are cured.

Motivated by the oropharynx cancer data set, but not limited to it, this model has a broad application in biomedical studies. For example, physicians will rely on a threshold in a biomarker or combination of biomarker signatures to guide the choice of therapy for an individual patient. The therapy targeting a specific biomarker generally will work effectively for patients when that biomarker is highly expressed, and thus, it is convenient to find a threshold in the biomarker such that therapy should only be given to those patients with the biomarker levels exceeding the threshold. In general, better characterizing cancers at the molecular level will lead to more efficient treatment, and methodology to improve estimates of a threshold point will help this characterization.

Several authors considered a Cox model with an unknown threshold in the covariate. Liang *et al.* [1] and Pons [2] proposed approaches that can be used to study the model as described in the motivating example if no cure fraction is present in the data, where the influence of a covariate z_1 (e.g., gender)

[a]*Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.*
[b]*Biometrics Research, Merck Research Lab, Rahway, NJ, U.S.A.*
[c]*University of Michigan Comprehensive Cancer Center, Ann Arbor, MI, U.S.A.*
**Correspondence to: Lili Zhao, Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.*
[†]*E-mail: zhaolili@umich.edu*

jumps at a certain threshold of z_2 (e.g., age). Luo and Boyett [3] studied a model where a constant is added to the regression on a covariate z_1 after a threshold in z_2. In this model, the baseline hazard changes after the threshold. Jensen and Lutkebohmert [4] and Kosorok and Song [5] considered a Cox-type regression model with a piecewise linear functional form of the covariates. However, the aforementioned models are not appropriate when a cured fraction is present in the population.

Othus *et al.* [6] estimated a threshold in a covariate in the cure model setting. They assume that there is a threshold in a covariate z_2, where a sudden jump or fall occurs in the hazard value or cure probability. But their model is restricted to a simple binomial–exponential mixture model, in which a binomial model is used to estimate the cure rate and exponential distribution is used for the conditional survival. In this paper, we extend the threshold detection problem to a more general cure model in which a logistic regression is used to evaluate the effect of covariates on the cure rate and a standard proportional hazard model is used for the conditional survival. This mixture cure model (without threshold detection) has been studied by many authors [7–9], and they use expectation maximization type algorithms to compute the maximum likelihood estimates. We build on this previous research to implement a Bayesian estimation method for the Cox proportional hazards cure model and extend it to allow a threshold in the regression coefficients. We will show that applying Bayesian methods in the mixture cure rate model is straightforward. Using a data augmentation scheme, the latent cure indicators are updated. As we demonstrate later, conditional on these indicators, the cure model reduces to a standard Cox model and a standard logistic regression model. Furthermore, our ultimate goal of detecting a threshold is simplified to a threshold problem in the regular Cox or logistic regression model.

The rest of the article is organized as follows: Section 2 outlines the model; Section 3 presents the Bayesian estimation and evaluation of model fit; Section 4 provides simulation studies; the analysis of the oropharynx cancer data is presented in Section 5; the paper ends with a discussion in Section 6.

2. Model description

The survival time, T, is assumed to be $T = vT^* + (1-v)\infty$, where v is an indicator of whether a subject will eventually ($v = 1$) or never ($v = 0$) experience the event, and T^* denotes the failure time if the subject is not cured. $S(t|v = 1)$ is the conditional survival function for patients who will experience the event, often called the latency distribution. The marginal survival function is $S(t) = c + (1-c)S(t|v = 1)$, where $c = P(v = 0)$.

We consider a Cox proportional hazard model in the latency part of the cure model. Similar to Liang *et al.* [1] and Pons [2], a threshold, τ, could be present in a time-independent covariate z_2, and the effect of z_1 changes after τ, specifically,

$$\Lambda(t|z_0, z_1, z_2, \tau; v = 1) = \Lambda_0(t)\exp\{\beta_0 z_0 + \beta_1 z_1 \mathrm{I}(z_2 \leqslant \tau) + \beta_2 z_1 \mathrm{I}(z_2 > \tau)\} \tag{1}$$

where $\mathrm{I}(z_2 \leqslant \tau)$ is a generic indicator function, which takes value of 1 if $z_2 \leqslant \tau$. β_1 represents the effect of z_1 for $z_2 \leqslant \tau$, and β_2 represents the effect of z_1 for $z_2 > \tau$. z_0 is a vector of baseline covariates. Let $\beta = (\beta_0, \beta_1, \beta_2)$ and $\tilde{z}(\tau) = (z_0, z_1\mathrm{I}(z_2 \leqslant \tau), z_1\mathrm{I}(z_2 > \tau))$, and z_0 could include z_2 but is distinct from z_1. $\Lambda_0(t)$ is an unspecified cumulative baseline hazard function. $S(t|\tilde{z}(\tau), v = 1) = \exp\{-\Lambda(t|\tilde{z}(\tau), v = 1)\}$.

A cure fraction c is modeled by a logistic regression or a probit model. In logistic regression, $c(x) = P(v = 0|x) = \frac{\exp(\gamma x)}{1+\exp(\gamma x)}$, and the vector of covariates x includes the intercept. In a probit model, $c(x) = \Phi(\gamma x)$, and Φ is the CDF of a standard normal distribution. These models can be extended to include a threshold in a covariate similar to that in the latency model.

3. Bayesian estimation and model selection

In practice, we observe $(t_i, \delta_i, x_i, z_i)$, and $i = 1, \cdots, n$, where t_i, denotes the observed survival time for the ith patient, δ_i is 0 if t_i is censored and 1 otherwise, and $z_i = (z_{0i}, z_{1i}, z_{2i})$ is a vector of covariates that may associate with the risk of experiencing the event, and x_i is a vector of covariates associated with the chance of cure. x_i and z_i could be identical.

It follows that $v_i = 1$ if $\delta_i = 1$, but if $\delta_i = 0$, v_i is unknown, and it can be 1 or 0. The probability that a censored patient will eventually experience the event is given by

$$p_{v_i} = P(v_i = 1|x_i, z_i, T > t_i) = \frac{(1 - c(x_i))S(t_i|z_i, v_i = 1)}{(1 - c(x_i))S(t_i|z_i, v_i = 1) + c(x_i)} \tag{2}$$

In Bayesian sampling, a data augmentation algorithm, described by Smith and Roberts [10] and Diebolt and Roberts [11], arises naturally for estimating the missing data, which in this case is an indicator of whether the patient is cured or uncured. A vector of $\boldsymbol{v} = (v_1, \cdots, v_n)$ is introduced as a vector of latent Bernoulli random variables. For patient i with $\delta_i = 0$, $v_i \sim \text{Bernoulli}(p_{v_i})$. Conditional on the vector of $(v_1, \cdots, v_n)$, the model reduces to the standard Cox model for patients with $v_i = 1$. In the incidence part, the model reduces to the standard logistic regression model, in which the vector of $(v_1, \cdots, v_n)$ is regressed on covariates $(x_1, \cdots, x_n)$.

3.1. Bayesian inference

We formulate the standard Cox model using counting processes [12]. Let $N_i(t)$ be the number of events that occurred up to time t. Let $\lambda_i(t)$ be the intensity function of $N_i(t)$, that is, $E(dN_i(t)|\mathcal{F}_{-t}) = \lambda_i(t)dt$, where $dN_i(t)$ is the increment of N_i over the small time interval $[t, t + dt)$. If subject i experiences the event during this interval, $dN_i(t)$ will take the value 1; otherwise, $dN_i(t)$ is 0. $\mathcal{F}_{-t}$ represents the available data just before time t. Then the proportional hazards model takes the form $\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta\tilde{z}_i(\tau)\}$, where $Y_i(t)$ is 1 if subject i is under observation at time t and 0 otherwise.

The counting process increments $dN_i(t)$ in the time interval $[t, t + dt)$ are assumed to be independent Poisson random variables with means $\lambda_i(t)dt = Y_i(t)\exp\{\beta\tilde{z}_i(\tau)\}d\Lambda_0(t)$, where $d\Lambda_0(t)$ is the increment in the cumulative baseline hazard function during the time interval $[t, t + dt)$. Given $\boldsymbol{v}$, the formulation for the latency applies only to patients with $v_i = 1$.

The time intervals are constructed based on the ordered distinct event times, $\{s_j; j = 1, \cdots, J\}$, where J is the total number of distinct times, s_J is the maximum observed event time, and s_{J+1} is infinity. The observed data D are assumed to be available within these intervals, such that $D = \{\mathcal{R}_j, \mathcal{D}_j, z_i, j = 1, \cdots, J; i = 1, \cdots, n\}$, where $\mathcal{R}_j$ is the risk set and $\mathcal{D}_j$ is the event set in interval $[s_j, s_{j+1})$. Let $\boldsymbol{\theta} = (\beta, d\Lambda_0(j), j = 1, \cdots, J)$, and the likelihood function for the aforementioned model is

$$L(\boldsymbol{\theta}, \tau; \boldsymbol{v} = 1) = \prod_{j=1}^{J} G_j$$

where

$$G_j = \exp\left\{-\sum_{i\in\mathcal{R}_j} d\Lambda_0(j)\exp\{\beta\tilde{z}_i(\tau)\}\right\}\prod_{i\in\mathcal{D}_j} d\Lambda_0(j)\exp\{\beta\tilde{z}_i(\tau)\} \tag{3}$$

The gamma process is used as a prior for the cumulative baseline hazard function Λ_0 [13]. That is, $\Lambda_0 \sim \mathcal{GP}(c_0\Lambda_0^*, c_0)$, where Λ_0^* is often assumed to be a known parametric function. For example, $\Lambda_0^* = \eta y^{k_0}$ corresponds to the Weibull distribution, and c_0 represents the degree of confidence in this prior guess. The prior distribution for $\beta \sim \text{MVN}(\mu_0, \Sigma_0)$, where μ_0 and Σ_0 are pre-specified.

In this study, τ is a parameter to be estimated from the data. Let $z_{2_{(1)}} < z_{2_{(2)}} < \cdots, < z_{2_{(K-1)}}$ be distinct ordered values of $z_{2i}, i = 1, \cdots, n$, and $z_{2_{(K)}}$ is the largest value of z_2. We propose to sample τ in two steps. The first step follows the work of Carlin *et al.* [14] and Lange *et al.* [15], in which the threshold is treated as a discrete variable in an application to simple regression models and Poisson processes. We first sample τ from a categorical distribution taking the value $z_{2_{(k)}}$ with probability π_k; that is,

$$\tau \sim \text{Multinomial}(1, (\pi_1, \cdots, \pi_k, \cdots, \pi_{K-1})) \tag{4}$$

$$\pi_k = \frac{L\left(\boldsymbol{\theta}, z_{2_{(k)}}; \boldsymbol{v} = 1\right)}{\sum_{k=1}^{K-1} L\left(\boldsymbol{\theta}, z_{2_{(k)}}; \boldsymbol{v} = 1\right)}$$

Given the intervals, the distribution of τ is assumed to be continuous with a uniform distribution, $z_{2_{(1)}} < \tau < z_{2_{(K)}}$. Thus, having obtained a draw of $z_{2_{(k)}}$ in step 1, we will sample τ from Uniform $[z_{2_{(k)}}, z_{2_{(k+1)}}]$ in the second step, which will result in a continuous posterior distribution of τ.

As an alternative to the aforementioned Gibbs sampling approach, we can directly consider τ as a continuous variable, and the conditional posterior density of τ can be written as

$$\tau \propto \prod_{j=1}^{J} G_j \mathrm{I}(\min(z_2) < \tau < \max(z_2))$$

We use the adaptive Metropolis algorithm [16] to sample τ. Specifically, we consider the proposal distribution given at iteration l by

$$Q^{(l)}(\tau^*, .) = (1 - B)\mathrm{N}\left(\tau^*, 2.38^2\sigma_\tau^{(l)}\right) + B\mathrm{N}(\tau^*, 0.1^2)$$

where τ^* is a candidate value for τ simulated from proposal $Q^{(l)}(\tau^*, .)$. $\sigma_\tau^{(l)}$ is the empirical estimate of the variance of the target distribution based on the entire history up to lth iteration. As suggested in Roberts and Rosenthal [16], we take B to be 0.05.

When no threshold is present in the model, τ is not identifiable [17]. It is possible to estimate the existence of a threshold in a mixture model as described by Skates *et al.* [18], in which reversible jump Markov chain Monte Carlo (MCMC) was used to move between a linear model with no threshold and a model with a threshold. But in our more complicated setting of a Cox model with a cured fraction, we have considered two other strategies to evaluate the presence of a threshold. First, we use model selection criteria to compare the models with and without a threshold. Second, we constrain τ in the range $(\min(z_2), \max(z_2))$, and let the non-identifiability be reflected in the posterior distribution of τ and $\beta_2 - \beta_1$. Our goal is to identify a sharp estimate of τ when the data clearly indicate a threshold. When the data do not clearly indicate a threshold, estimates of τ would have large uncertainty, and the contrast parameter of $\beta_2 - \beta_1$ would be close to 0.

We can thus carry out the following hybrid Gibbs sampling scheme:

1. Sample $v_i \sim \mathrm{Bernouli}(p_{v_i})$ for patients with $\boldsymbol{\delta} = 0$, where p_{v_i} is defined as in (2).

Steps 2–4 are applied for patients with $v_i = 1$.

2. Sample from

$$P(\beta|\tau, \Lambda_0, D) \propto \prod_{j=1}^{J} G_j \exp\left\{-\frac{1}{2}(\beta - \mu_0)\Sigma_0^{-1}(\beta - \mu_0)\right\}$$

   using random-walk Metropolis algorithm as developed by Haario *et al.* [19].
3. Sample $\Lambda_0(j), j = 1, \cdots, J$ as

$$d\Lambda_0(j) \sim \mathcal{G}\left(c_0\eta(s_{j+1} - s_j) + d_j, c_0 + \sum_{i \in \mathcal{R}_j} \exp\{\beta\tilde{z}_i(\tau)\}\right)$$

   where d_j is the number of events in $[s_j, s_{j+1})$.
4. Sample τ as defined in (4), or sample τ using the adaptive Metropolis algorithm as developed by Roberts and Rosenthal [16].
5. Sample γ using random-walk Metropolis algorithm [19] in the logistic regression model. For a probit model, the truncated normal sampling approach proposed by Albert and Chib [20] can be used. A multivariate normal prior was used for γ.
6. Update $S(t_i|\tilde{z}_i(\tau), v_i = 1) = \exp\left\{-\sum_{j=1}^{J} Y_i(s_j)d\Lambda_0(j)\exp\{\beta\tilde{z}_i(\tau)\}\right\}$ and $c(x_i) = c_i = \frac{\exp(\gamma x_i)}{1+\exp(\gamma x_i)}$, and update p_{v_i} as a function of $c(x_i)$ and $S(t_i|\tilde{z}_i(\tau), v_i = 1)$. If there is no covariate in the incidence model, we can estimate the cure rate c by a logistic regression model with just an intercept or simply averaging over the indicator variables $v_1, \cdots, v_n$.

3.2. Model selection

We computed two Bayesian model comparison criterions for selecting the best model. To alleviate the concern that the standard deviance information criterion (DIC) measure [21] does not properly reflect the correct effective number of parameters in mixture models, Celeux [22] recommended a modified DIC, termed DIC_3, which estimates $D(\mathrm{E}[\boldsymbol{\theta}|y])$ using the posterior mean of the observed likelihood averaged across the cured and uncured subjects. Specifically,

$$\mathrm{DIC}_3 = -\frac{4}{m}\sum_{l=1}^{m}\left\{\sum_{i=1}^{n}\log f^{(l)}(y_i)\right\} + 2\sum_{i=1}^{n}\log\left\{\frac{1}{m}\sum_{l=1}^{m} f^{(l)}(y_i)\right\}$$

$f^{(l)}(y_i)$ is approximated by

$$\left\{\left(1-c_i^{(l)}\right)\sum_{j=1}^{J} dN_i(s_j)d\Lambda_0^{(l)}(j)\exp\left\{\beta^{(l)}\tilde{z}_i(\tau^{(l)})\right\}S^{(l)}(t_i|\tilde{z}_i(\tau), v_i=1)\right\}^{\delta_i}$$
$$\times\left\{c_i^{(l)} + \left(1-c_i^{(l)}\right)S^{(l)}(t_i|\tilde{z}_i(\tau), v_i=1)\right\}^{1-\delta_i}$$

where m is the number of draws of the posterior distribution. $c_i^{(l)} = \frac{\exp(\gamma^{(l)}x_i)}{1+\exp(\gamma^{(l)}x_i)}$, $d\Lambda_0^{(l)}(j)$, $\beta^{(l)}$, and $S^{(l)}(t_i|\tilde{z}_i(\tau), v_i=1)$ are the values of the parameters for the lth draw.

The log-pseudo-marginal likelihood (LPML) [23] is a cross-validated leave-one-out measure of a model's ability to predict the data. It is valid for small and large samples and does not suffer from a heuristic justification based on large sample normality. LPML is defined based on the conditional predictive ordinate (CPO) statistic for the ith observation, and CPO_i is given by $\mathrm{CPO}_i = f(D_i|D^{(-i)})$, where D_i denote the ith observation, and $D^{(-i)}$ denote the data with ith observation deleted. The log of the product of the CPO statistics under a given model is the LPML statistic for that model, $\mathrm{LPML} = \sum_{i=1}^{n}\log \mathrm{CPO}_i$. The model with larger LPML is preferred. A Monte Carlo approximation of CPO_i is given by Chen *et al.* [24]: $\mathrm{LPML} = -\sum_{i=1}^{n}\log\left\{\frac{1}{m}\sum_{l=1}^{m} 1/f^{(l)}(y_i)\right\}$.

4. Simulations

Simulation studies were conducted to evaluate the proposed approach. All simulations consist of 1000 experimental replications, each with sample size of $n = 200$ or $n = 400$. Survival times, T, are generated from a logistic–exponential mixture model, where $c(z) = (1+\exp(-\gamma_1-\gamma_2 z_1-\gamma_3 z_2))^{-1}$, and $S(t|v=1;z) = \exp\{-\exp(\beta_0 z_2+\beta_1 z_1 \mathrm{I}(z_2\leqslant\tau)+\beta_2 z_1\mathrm{I}(z_2>\tau))t\}$. In this formulation, the baseline hazard function is constant with a rate of 1. z_1 and z_2 are fixed by design. $z_1 = 0.5$ for half of the sample size and $z_1 = -0.5$ for the other half, and the covariate z_2 is generated from uniform [0, 10]. The τ is assumed to be 3, 5, 7, or 15 in which no threshold is present. Each subject is followed up until at most time = 5. Censoring times C are generated from an exponential distribution with censoring rate of 4.5. The data for each observation are (t, δ, z_1, z_2), where $t = \min(T, C, 5)$. With the choices of the parameters listed in Table I, the expected censoring proportion including those cured is around 0.54, and the observed cure rate is around 0.43.

The models were implemented in **R**. A multivariate t-distribution with a degree of freedom of 3 was used as the proposal density in the random-walk Metropolis algorithm in sampling β and γ. The proposal density centered at the previous value, and the covariance was adaptive as developed by Haario [19], which uses the empirical covariance from an extended burn-in period. We proposed two algorithms to estimate the threshold. We found that the adaptive Metropolis algorithm in step 4 perform better than the discrete algorithm. For the rest of this article, the adaptive Metropolis algorithm in step 4 is used. Patients who survive after the last observed survival time are considered as cured in the estimation procedure. We observed that the chain mixes well. The priors are quite vague relative to the likelihood: a vector of zeros is the prior mean of β and γ; Σ_0 is the prior covariance matrix with 100 on the diagonal for β and γ. In the gamma process prior, Λ_0^* is assumed to have an exponential distribution ($k_0 = 1$) with $\eta = 0.1$ and $c_0 = 0.1$, namely $d\Lambda_0(j) \sim \mathcal{G}(0.1\times 0.1(s_{j+1}-s_j), 0.1)$. The value of 0.1 for η underestimates the true value of η, but a small c_0 of 0.1 gives large uncertainty about this η.

With a burn-in of 5000 iterations, an additional 10,000 iterations were used for inference. Results from Table I indicate that the proposed model accurately estimates the true values of all the parameters

Table I. List of parameters used in simulation and performance statistics.

		Mean (ESE)[a]	SSE[b]	MSE	CP[c]	Mean (ESE)	SSE	MSE	CP
Parameter	True		$n = 200$				$n = 400$		
τ	5	5.01 (0.98)	0.72	0.52	95	5.00 (0.39)	0.39	0.15	96
β_0	0.05	0.05 (0.04)	0.04	0.002	95	0.05 (0.03)	0.03	0.001	94
β_1	−1	−0.98 (0.50)	0.51	0.26	95	−0.97 (0.29)	0.29	0.08	95
β_2	1	1.02 (0.40)	0.40	0.16	94	1.00 (0.24)	0.23	0.05	95
γ_1	0.2	0.21 (0.35)	0.36	0.13	95	0.22 (0.24)	0.25	0.06	93
γ_2	0.5	0.57 (0.35)	0.36	0.13	94	0.52 (0.24)	0.24	0.06	93
γ_3	−0.1	−0.11 (0.06)	0.06	0.004	95	−0.10 (0.04)	0.04	0.002	93
$\beta_2 - \beta_1$	2	2.00 (0.64)	0.63	0.40	96	1.97 (0.37)	0.37	0.14	97
τ	7	6.65 (1.03)	0.80	0.77	96	6.96 (0.39)	0.37	0.14	95
β_0	0.05	0.05 (0.04)	0.04	0.002	95	0.05 (0.03)	0.03	0.002	95
β_1	−1	−1.00 (0.44)	0.44	0.19	96	−0.97 (0.23)	0.24	0.06	94
β_2	1	0.98 (0.53)	0.60	0.36	94	1.01 (0.31)	0.31	0.10	95
γ_1	0.2	0.21 (0.35)	0.36	0.13	95	0.22 (0.24)	0.25	0.06	93
γ_2	0.5	0.58 (0.35)	0.36	0.14	94	0.53 (0.24)	0.24	0.06	94
γ_3	−0.1	−0.11 (0.06)	0.06	0.004	95	−0.10 (0.04)	0.04	0.001	94
$\beta_2 - \beta_1$	2	1.98 (0.70)	0.75	0.57	96	1.98 (0.38)	0.39	0.15	96
τ	3	3.56 (1.33)	1.00	1.32	96	3.15 (0.58)	0.59	0.37	95
β_0	0.05	0.05 (0.04)	0.04	0.002	96	0.05 (0.03)	0.03	0.001	94
β_1	−1	−0.89 (0.75)	0.84	0.71	93	−0.93 (0.41)	0.41	0.17	94
β_2	1	1.03 (0.41)	0.43	0.19	96	0.99 (0.22)	0.25	0.06	95
γ_1	0.2	0.20 (0.35)	0.35	0.12	95	0.21 (0.24)	0.25	0.06	93
γ_2	0.5	0.55 (0.36)	0.37	0.14	94	0.50 (0.24)	0.24	0.06	94
γ_3	−0.1	−0.10 (0.06)	0.06	0.004	95	−0.10 (0.04)	0.04	0.00	93
$\beta_2 - \beta_1$	2	1.91 (0.86)	0.94	0.89	97	1.93 (0.46)	0.49	0.25	95
τ	No	4.89 (2.81)	1.36	—	—	4.89 (2.86)	1.40	—	—
β_0	0.05	0.05 (0.04)	0.04	0.002	95	0.05 (0.03)	0.03	0.001	94
β_1	−1	−0.97 (1.01)	1.03	1.06	97	−1.03 (0.79)	0.79	0.62	98
β_2	−1	−0.99 (0.85)	0.94	0.89	98	−1.01 (0.63)	0.81	0.66	97
γ_1	0.2	0.21 (0.35)	0.35	0.13	94	0.22 (0.24)	0.25	0.06	94
γ_2	0.5	0.58 (0.35)	0.36	0.14	94	0.55 (0.24)	0.24	0.06	94
γ_3	−0.1	−0.11 (0.06)	0.06	0.004	95	−0.10 (0.04)	0.04	0.002	93
$\beta_2 - \beta_1$	0	−0.02 (1.45)	1.42	2.01	95	0.02 (1.13)	1.15	1.31	96

MSE, mean square error.

[a]Average of the posterior means over 1000 data sets (average of the posterior standard deviations over the 1000 data sets).

[b]Standard deviation of the posterior means across 1000 data sets.

[c]Coverage of the 95 percentage highest posterior density interval.

Table II. Percent of times threshold model is chosen over non-threshold model.

Selection criterion	n	$\tau = 3$	$\tau = 5$	$\tau = 7$	No threshold
DIC_3[a]	200	75	88	86	8
	400	97	100	100	6
LPML	200	75	80	80	50
	400	86	93	93	50

DIC, deviance information criterion; LPML, log-pseudo-marginal likelihood.

[a]In DIC_3, threshold model is selected if its DIC_3 is less than the non-threshold model by more than 3.

regardless of different values of τ. The accuracy is further improved when sample size is increased to $n = 400$, as evidenced by significantly reduced standard deviations and mean squared error (MSE) given about 95% coverage probability.

In the last scenario where no threshold is present in z_2, the mean of τ is around 5, which is in the middle of the range z_2 from 0 to 10, but with large standard deviations (ESE $\approx$ 3). The reason is that τ is constrained to be in the range of z_2 in each Gibbs sampling step, thereby leading to an average z_2

to be the estimate of τ, and the standard deviation is increased when n is increased to 400. Furthermore, the hypothesis of no threshold is strengthened by testing the contrast of $\beta_2 - \beta_1$ (the point estimates are very close to 0: -0.02 with $n = 200$ and 0.02 with $n = 400$). From 1000 experimental replications, the 95% highest posterior density (HPD) intervals of the contrast include 0 for about 95% of the times.

As a rule of thumb, if two models differ in DIC by more than 3, the one with smaller DIC is preferred as the best fitting [21]. Based on both DIC_3 and LPML, threshold models are preferred to the models that ignore the threshold (called non-threshold model) when a threshold is truly present. As shown in Table II, when $\tau = 5$, 88% (based on DIC_3) and 80%(based on LPML) of the time, the threshold model is chosen over the non-threshold model. In all cases studied, the correct models are chosen more frequently when n is increased to 400. When there is truly no threshold, DIC_3 tends to prefer the simpler non-threshold model, evidenced by the non-threshold model chosen over the threshold model 92% (100–8%) of the time. In contrast, LPML seems to have no penalty for more complicated models; half of the time, the threshold model is chosen although no threshold is present. In addition, using 3 as a threshold for DIC_3 comparisons seems to yield reasonable types I and II errors.

Normal priors for β and γ are routinely used in the regression models. In this application, we adopted very vague priors for these parameters (relative to the likelihood) in the aforementioned simulation studies and the oropharynx cancer example in the next section. However, the posterior distribution of τ could be sensitive to the gamma process prior. This prior consists of an initial estimate of the cumulative baseline hazard function Λ_0^* and a precision c_0. In reality, an exponential distribution of Λ_0^* is used mostly for convenience. It is easy to have a prior guess on the average event rate and assume it to be constant in the study period. Therefore, it is important to evaluate the robustness of the exponential distribution in the proposed model. For this purpose, we repeated the aforementioned simulation when $\tau = 5$ and generated the baseline hazard function from three Weibull distributions: scenario I (Weibull(0.9, 1.33)), scenario II (Weibull(1.1, 0.8)), and scenario III (Weibull(1, 1)), where a is a scale parameter and b is a shape parameter in Weibull(a, b). In all the three scenarios, the average rate is 1. We then applied the proposed model using an exponential distribution for Λ_0^* with a rate equal to 1 (the same as the true rate) or 0.1 (much lower than the true rate). We also varied the degree of confidence in the prior guess by considering $c_0 = 0.01, 0.1, 5$, and 10.

As shown in Table III, the exponential prior (when $k_0 = 1$) works well even when the hazard function is not exponential, and using the Weibull distribution did not improve the performance even when the parameters in the Weibull prior are the same as the truth. The independent assumption between the disjoint intervals in the posterior inference makes the parametric form of the hazard function less important and the rate per interval more important. As expected, when η is incorrectly specified (e.g., $\eta = 0.1$), a large precision, c_0, resulted in considerably increased bias, ESE, and MSE.

5. Oropharynx cancer example

We now elaborate on and analyze the data from the motivating example in Section 1. The data were collected from 220 patients with oropharynx cancer enrolled in the University of Michigan Head and Neck Cancer Specialized Programs of Research Excellence during the years 2003 to 2008. Of the 220 patients, 84 died, 55 of whom died from oral cavity cancer, and the remaining 136 were alive at the end of follow-up. Of the 220 patients, 36 are female and 184 are male; 11 patients had stage II, 37 had stage III, and 172 had stage IV. The mean age is 58 years (range from 22 to 86 years). A K-M survival curve (Figure 1) for the whole data set has a level region beyond about 60 months, which may indicate the appropriateness of a cure model. Patients who survive after the last observed survival time are considered as cured in our mixture model, which effectively eliminates the problem of lack of identifiability [25].

The priors for β and γ are the same as in the simulation studies. They are very vague relative to the likelihood. To construct a reasonable prior for the baseline hazard, we fit our proposed model using data from all head and neck cancer patients excluding the oropharynx cancer patients ($n = 256$), and we obtained an estimate of 0.1 for the baseline hazard rate of η assuming that the hazard is constant over time (a similar strategy was used in Ibrahim *et al.* [26] to construct a prior for the baseline hazard rate). We then set c_0 to be 0.01 to reflect the uncertainty of our estimate of the hazard rate in the oropharynx population. The prior for the other parameters is the same as in the simulation studies. With a burn-in of 20,000 iterations, an additional 20,000 iterations were used for inference. We observed that the chain mixes well, and the results are robust to different choices of the initial values.

Table III. Performance of τ under different gamma process priors.

Scenario	Prior[a] η	k_0	c_0	Mean (bias)[b]	ESE[c]	MSE	CP[d]
I	0.1	1	0.01	4.97 (0.45)	0.90	0.48	95
	1	1	0.01	5.00 (0.04)	0.90	0.49	95
	0.1	1	0.1	4.99 (0.11)	0.91	0.47	95
	1	1	0.1	5.00 (.001)	0.90	0.50	95
	0.1	1	5	4.93 (1.47)	1.08	0.50	97
	1	1	5	5.02 (0.36)	0.90	0.47	96
	0.1	1	10	4.84 (3.16)	1.13	0.59	96
	1	1	10	5.00 (.008)	0.91	0.44	96
	∗0.9	1.33	0.1	4.99 (0.12)	0.88	0.46	96
	∗0.9	1.33	5	5.00 (0.09)	0.79	0.52	94
II	0.1 1		0.01	4.99 (0.11)	1.06	0.55	96
	1	1	0.01	5.00 (0.04)	1.05	0.54	96
	0.1	1	0.1	1.99 (0.23)	1.07	0.50	96
	1	1	0.1	5.00 (0.09)	1.05	0.52	96
	0.1	1	5	4.89 (2.20)	1.15	0.54	97
	1	1	5	5.02 (0.34)	0.98	0.53	96
	0.1	1	10	4.83 (3.42)	1.17	0.69	97
	1	1	10	5.01 (0.26)	0.95	0.50	95
	∗1.1	0.8	0.1	4.99 (0.24)	1.05	0.51	96
	∗1.1	0.8	5	5.01 (0.17)	1.05	0.58	97
III	0.1	1	0.01	4.99 (0.21)	0.97	0.49	96
	∗1	1	0.01	5.00 (0.08)	0.97	0.54	96
	0.1	1	0.1	5.01 (0.07)	0.98	0.52	95
	1	1	0.1	4.99 (0.23)	0.98	0.51	95
	0.1	1	5	4.92 (1.75)	1.11	0.53	97
	∗1	1	5	5.01 (0.28)	0.93	0.53	95
	0.1	1	10	4.82 (3.43)	1.14	0.60	97
	∗1	1	10	5.00 (.007)	0.92	0.50	95

MSE, mean squared error.

[a]Scenarios denoted by ∗ are those when the prior matches the true distribution.

[b]Average of the posterior means over 1000 data sets (bias is defined as $|(\text{Mean} - 5)/5 \times 100|$).

[c]Average of the posterior standard deviation over 1000 data sets.

[d]Coverage of the 95 percentage highest posterior density interval.

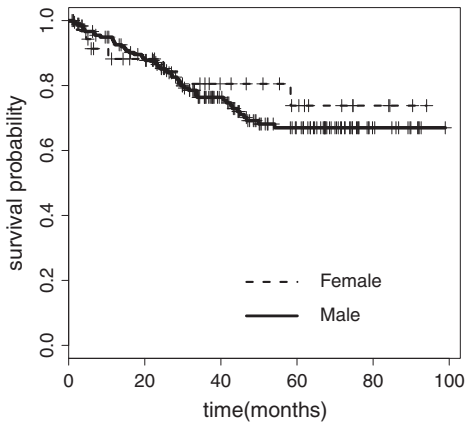

Figure 1. Kaplan–Meier disease-specific survival plots for all ages.

As shown in Table IV, there is no difference in disease-specific survival between women and men when age is not considered (see model 1). However, a significant gender effect is revealed when the analysis is conditional on age. Model 2 indicates an interaction between gender and age in the latency,

Table IV. Parameter estimates and performance statistics for the oropharynx cancer example.

	Covariate	Mean	SE	95% HPD
Model 1: Mixture cure model				
Latency	Sex (F vs M)	−0.16	0.54	(−1.23, 0.85)
	Age	−0.003	0.02	(−0.04, 0.03)
	Stage	−0.13	0.42	(−0.92, 0.74)
Incidence	Sex (F vs M)	0.28	0.56	(−0.83, 1.45)
	Age	−0.008	0.02	(−0.04, 0.03)
	Stage	−0.34	0.49	(−1.24, 0.68)
Model 2: Mixture cure model with an interaction				
Latency	Sex (F vs M)	0.002	0.56	(−1.02, 1.08)
	Age	0.001	0.02	(−0.03, 0.03)
	Stage	0.13	0.44	(−0.71, 1.03)
	Sex × Age	−0.18	0.05	(−0.28, −0.09)
Incidence	Sex (F vs M)	0.28	0.52	(−0.78, 1.30)
	Age	−0.01	0.02	(−0.04, 0.03)
	Stage	−0.20	0.45	(−1.10, 0.66)
Model 3: Mixture cure model with a threshold				
Latency	Threshold (τ)	65	7	(51, 75)
	Sex (age $\leqslant \tau$) (F vs M)	1.50	0.73	(0.04, 2.91)
	Sex (age $> \tau$) (F vs M)	−2.92	1.38	(−5.45, −0.17)
	Age	0.001	0.02	(−0.03, 0.04)
	Stage	0.03	0.45	(−0.78, 0.95)
Incidence	Sex (F vs M)	0.35	0.52	(−0.61, 1.36)
	Age	−0.008	0.02	(−0.04, 0.02)
	Stage	−0.26	0.44	(−1.04, 0.67)

SE, standard error; HPD, highest posterior density.

Table V. Model comparisons.

Models	DIC_3 (p_D)	LPML
1	795 (66)	−470
2	780 (65)	−464
3	783 (66)	−466
4	788 (66)	−467
5	794 (66)	−475
6	781 (66)	−464
7	783 (66)	−465

DIC, deviance information criterion; LPML, log-pseudo-marginal likelihood.

and model 3 provides an estimate of a threshold in age such that the gender effect changes after the threshold. Both DIC and LPML favor models that considered the interaction between age and gender (Table V). Model 1 has significantly higher DIC and lower LPML than other models. Although model 2 has slightly lower DIC than model 3, the two models are considered to fit the data equally well using the cutoff of three (the choice of three is good in terms of the desired error rates as shown in the aforementioned simulation studies). The slightly higher LPML in model 2 seems to suggest that fitting the interaction using the continuous age is better than dichotomizing the age; however, a threshold can be important in clinical practice as emphasized in Section 1. We are willing to trade a little bit of goodness of fit of the model for a useful application, and this slight sacrifice is negligible in terms of the pseudo-Bayes factor for comparing model 2 to model 3, defined as $PBF_{23} = \exp(LPML_2 - LPML_3) = 2.7$ [27]. In model 3, the point estimate of τ is 65, and men have significantly better prognosis in disease-specific survival for patients younger than τ, and women have significantly better prognosis for patients older than τ (95% HPD interval of the contrast $\beta_2 - \beta_1$ is (1.8, 7.4)). Figure 2 presents the switched gender effect conditional on the point estimate of the threshold. This significant gender effect will not be detected when a threshold model is not considered.

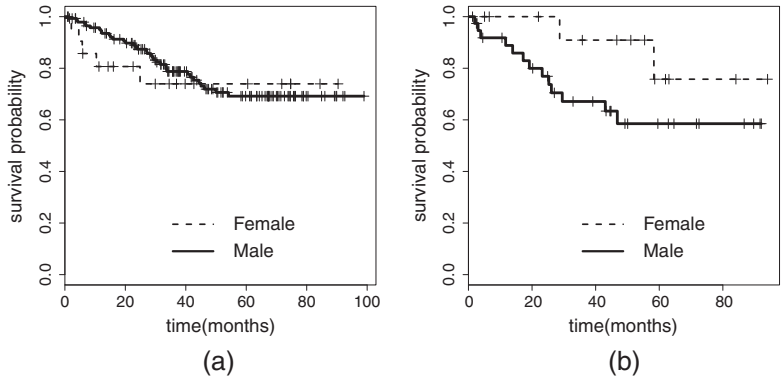

Figure 2. (a) Kaplan–Meier disease-specific survival plots for patients 65 years old or younger (point estimate of τ). (b) Kaplan–Meier disease-specific survival plots for patients older than 65 years.

To validate the presence of a cure fraction, we fitted the data using a regular Cox regression model with a threshold in age (called model 4), namely a model 3 without the incidence part. A considerably increased DIC_3 of 788, compared to model 3, indicates that the model with a cure fraction is a good choice. Without the cure fraction, we did not find a clear threshold in age, as evidenced by a large 95% HPD interval in τ ranging from 28 to 85. The gender effect before and after the τ is estimated with a large variance caused by the large uncertainty in estimating τ (results not shown).

In this application, we also considered three other models to evaluate the potential threshold of age in incidence. Model 5 assumes a threshold in the incidence rather than in the latency. A large DIC of 794 and a large 95% HPD interval of τ from ages 27 to 86 years supports the absence of a threshold in age in the incidence. Model 6 adds an additional interaction of age and gender in incidence to model 2. The interaction in the incidence is not significant, the 95% HPD interval of the interaction is $(-0.05, 0.13)$, and both the DIC and LPML are very close to model 2. Based on model 3, model 7 updates the parameters in the logistic regression given each realization of the threshold in latency. That is, model 7 assumes the same threshold in latency exists in incidence. Again, the gender effect is similar before and after the threshold in incidence, and both the DIC and LPML are very close to model 3. The two stable plateaus in the K-M curves are not statistically different in Figure 2(a, b), which seems to further confirm that no threshold is present in the incidence. The results for models 6 and 7 were not shown as the parameter estimations are similar to models 2 and 3, respectively.

6. Discussion

In this paper, we develop a Bayesian approach to estimate a Cox proportional hazards cure model and extend the Cox model to allow a threshold in the regression coefficient. Personalized therapy is the future of oncology drug development. Dichotomizing a continuous biomarker is a common practice in clinical research because it facilitates a decision to be made about which therapy to give and is easy to include in protocols. Compared to other methods in finding the optimal threshold for categorizing continuous variable by maximizing some likelihood function or test statistic [28, 29], our method has the advantages of (i) taking into account the sampling variation in estimating the threshold, as well as other parameters in the model that depend on the variable threshold; (ii) obtaining a distribution of the threshold; (iii) adjusting for other prognostic variables when estimating the threshold; (iv) directly testing the absence or presence of a threshold; and (v) evaluating the assumption of a sudden jump of a covariate at the threshold by comparing to a model with an interaction term using model selection criterions.

The introduction of the latent Bernoulli cure indicators greatly facilitates the MCMC algorithm. Given the indicators, the latency and incidence can be evaluated separately using standard methods, and the threshold detection problem reverts to a problem in the Cox model. Chen *et al.* [30] noted that Bayesian inference for a mixture cure model requires proper priors to avoid the possibility of improper posterior distributions. We avoid this issue by using very mildly informative but proper priors.

In this study, we are interested in a threshold in the latency. We found that the MCMC algorithm had high autocorrelation and slow convergence when we tried to estimate a threshold in both latency and incidence for the same covariate. If the threshold in the latency model is your primary interest, you can

simply update the parameters in the logistic regression given each realization of the threshold in latency, as describe in model 5 in the oropharynx cancer example.

In this paper, we have used a mixture cure model in which covariate effects are separately considered for incidence and latency, and thus the threshold effect on covariates can also be considered separately on incidence and latency. We believe that a covariate that is important for long-term incidence may not be important for latency and vice versa. We did not consider this threshold detection problem in the bounded cumulative hazard cure model [26, 30–33]. But it would be a nice alternative cure model if you believe that a covariate is equally important in both latency and incidence.

The parameter estimates (point or interval estimates) are calculated using the MCMC iterations from the proposed model rather than based on asymptotic approximation as in frequentist approaches. Moreover, estimates of the contrast $\beta_2 - \beta_1$ can be calculated easily to evaluate the existence of a threshold. This *ad hoc* way of testing the presence or absence of a threshold combined with the model selection criterions works well in our study.

In this application, we estimate a threshold in a continuous variable such that the effect of a dichotomized variable changes before and after the threshold. With some modification of the design matrix X, this method can be used to estimate a threshold in a covariate as described in Jensen and Lutkebohmert [4], where the covariate effect has a piecewise linear functional form. Extensions to multiple threshold detection using reversible jump MCMC [34] will be the subject of future research.

Acknowledgements

This work was partially supported by the University of Michigan Head and Neck Cancer Specialized Program of Research Excellence, grant number CA097248.

References

1. Liang K, Self S, Liu X. The Cox proportional hazards model with change point an epidemiologic application. *Biometrics* 1990; **79**:783–793. DOI: 10.2307/2532096.
2. Pons O. Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Annals of Statistics* 2003; **56**:442Ű–463. DOI: 10.1214/aos/1051027876.
3. Luo X, Boyett J. Estimation of a threshold parameter in Cox regression. *Communication of Statistical Theory Methods* 1997; **26**:2329–2346. DOI: 10.1093/biomet/79.3.531.
4. Jensen U, Lutkebohmert C. A Cox-type regression model with change-points in the covariates. *Lifetime Data Analysis* 2008; **14**:276–285. DOI: 10.1007/s10985-008-9083-3.
5. Kosorok M, Song R. Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics* 2007; **35**:957–989. DOI: 10.1214/009053606000001244.
6. Othus M, Li Y, Tiwari R. Change-point cure models with application to estimating the changepoint effect of age of diagnosis among prostate cancer patients. *Journal of Applied Statistics* 2012; **39**:901–911.
7. Kuk A, Chen C. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* 1992; **79**:531–541. DOI: 10.1093/biomet/79.3.531.
8. Peng Y, Dear K. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; **56**:237–243. DOI: 10.1111/j.0006-341X.2000.00237.x.
9. Sy J, Taylor J. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; **56**:227–236. DOI: 10.1111/j.0006-341X.2000.00227.x.
10. Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B* 1993; **55**:3–23.
11. Diebolt J, Roberts CP. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B* 1994; **56**:363–375.
12. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Annals of Statistics* 1982; **10**:1100–1120.
13. Kalbfleisch J. Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B* 1978; **40**:214–221.
14. Carlin BP, Gelfand AE, Smith AFM. Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society. Series C-Applied Statistics* 1992; **41**:389–405. DOI: 10.2307/2347570.
15. Lange N, Carlin BP, Gelfand AE. Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *Journal of the American Statistical Association* 1992; **87**:615–626. DOI: 10.2307/2290197.
16. Roberts GO, Rosenthal JS. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 2009; **18**:349–367. DOI: 10.1198/jcgs.2009.06134.
17. Andrews DWK. Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica* 2001; **69**:683–773. DOI: 10.1111/1468-0262.00210.

18. Skates SJ, Pauler DK, Jacobs IJ. Screening based on the risk of cancer calculation from Bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association* 2001; **96**:429–439. DOI: 10.1198/016214501753168145.
19. Haario H, Saksman S, Tamminen J. An adaptive Metropolis algorithm. *Bernoulli* 2001; **7**:223–242. DOI: 10.2307/3318737.
20. Albert J, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 1993; **88**:669–679. DOI: 10.2307/2290350.
21. Spiegelhalter DJ, Best NG, Carlin B, de Linde AV. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society* 2002; **64**:583–639. DOI: 10.1111/1467-9868.00353.
22. Celeux G, Forbes F, Robert C, Titterington D. Deviance information criteria for missing data models. *Bayesian Analysis* 2006; **1**:651–674. DOI: 10.1214/06-BA122.
23. Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. Springer: New York, 2001.
24. Chen MH, Shao Q, Ibrahim JG. *Monte Carlo Methods in Bayesian Computation*. Springer: New York, 2000.
25. Taylor JMG. Semi-parametric estimation in failure time mixture models. *Biometrics* 1995; **51**:899–907. DOI: 10.2307/2532991.
26. Ibrahim JG, Chen MH, MacEachern SN. Bayesian variable selection for proportional hazards models. *The Canadian Journal of Statistics* 1999; **27**:701–717. DOI: 10.2307/3316126.
27. Hanson TE. Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association* 2006; **101**:1548–1565. DOI: 10.1198/016214506000000384.
28. Clark GM, Dressler LG, Owens MA, Pound G, Oldaker T, McGuire W. Prediction of relapse or survival in patients with node-negative breast cancer by DNA flow cytometry. *New England Journal of Medicine* 1989; **320**:627.
29. Sigurdsson H, Baldetorp B, Borg A, Dalberg M, Ferno MM, Killander D, Olsson H. Flow cytometry in primary breast cancer: improving the prognostic value of the fraction of cells in the s-phase by optimal categorisation of cut-off levels. *Br. Journal of Cancer* 1990; **62**:786.
30. Chen M, Ibrahim J, Sinha D. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 1999; **94**:909–910. DOI: 10.2307/2670006.
31. Kim S, Chen MH, Dey DK, Gamerman D. Bayesian dynamic models for survival data with a cure fraction. *Lifetime Data Analysis* 2007; **13**:17–35. DOI: 10.1007/s10985-006-9028-7.
32. Tighiouart M. Modeling correlated time-varying covariate effects in a Cox-type regression model. *Journal of Modern Applied Statistics Methods* 2003; **2**:161–167.
33. Tsodikov A, Ibrahim JG, Yakovlev AY. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association* 2003; **98**:1063–1078. DOI: 10.1198/0162214503000000001007.
34. Green PJ. Reversible-jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**:711–732. DOI: 10.2307/2337340.