A Bayesian promotion time cure model with current status data

Pavithra Hariharan, P. G. Sankaran

Department of Statistics, Cochin University of Science and Technology, Cochin 682 022, Kerala, India Corresponding author email: pavithrahariharan97@gmail.com

Abstract

Analysis of lifetime data from epidemiological studies or destructive testing often involves current status censoring, wherein individuals are examined only once and their event status is recorded only at that specific time point. In practice, some of these individuals may never experience the event of interest, leading to current status data with a cured fraction. Cure models are used to estimate the proportion of non-susceptible individuals, the distribution of susceptible ones, and covariate effects. Motivated from a biological interpretation of cancer metastasis, promotion time cure model is a popular alternative to the mixture cure rate model for analysing such data. The current study is the first to put forth a Bayesian inference procedure for analysing current status data with a cure fraction, resorting to a promotion time cure model. An adaptive Metropolis-Hastings algorithm is utilised for posterior computation. Simulation studies prove our approach's efficiency, while analyses of lung tumor and breast cancer data illustrate its practical utility. This approach has the potential to improve clinical cure rates through the incorporation of prior knowledge regarding the disease dynamics and therapeutic options.

Keywords: Current status data, Cure rate, Promotion time cure model, Bayesian inference, Adaptive Metropolis-Hastings algorithm.

1. Introduction

Modelling of lifetime data becomes challenging due to the complexities arising from incomplete data, attributed to different forms of censoring. Let T>0 be a random variable quantifying the time taken for an event to take place. Current status censoring conceals T, but allows for observation of a single random monioring time U>0 and a random interval where T falls. Under a conventional current status data model, the data observable from each subject is (U,δ) , where $\delta=I$ ($T\leq U$) that takes value 1 for events within [0,U] and 0 for events within $(U,\infty]$. This category of data is identified as current status data (alternatively termed type I interval censored data), due to its nature of revealing only the present condition of the observed individual. These data might be preferred for destructive testing and epidemiological research, where repeated evaluations pose significant challenges due to practical, ethical, and logistical constraints. These constraints include limited resources, the invasive nature of testing, or the inherent design of the study, which makes continuous or frequent monitoring impractical (Groeneboom and Wellner 1992; Jewell and van der Laan 2003).

Examples comprise data on uterine fibroid development status at the ultrasound examination time arising out of a "Right from the Start" prospective cohort study held at three different states of United States (Laughlin et al. 2009), data on renal recovery status of acute kidney injured at the time of discharge from University of Michigan Hospital (Heung et al. 2012; Al-Mosawi and Lu 2022), and gonorrhea status data from a Nebraska Public Health Laboratory survey (Li et al. 2021). There has been substantial interest in the analysis of current-status data, see Diamond et al. (1986), Jewell and Shiboski (1990), Andersen and Ronn (1995), Li et al. (2024), Wang and Du (2024), and Wu et al. (2024). Some recent studies on Bayesian modelling utilising current status data belong to Cai et al. (2011), Das et al. (2024), and Paulon et al. (2024).

The aforementioned studies postulate that each person in the study encounters the event eventually, given adequate follow-up time. However, in today's healthcare system, wherein many fatal diseases are recoverable, there may exist certain individuals immune to the event, considered cured. This can cause the typical survival curve to plateau rather than decline to zero. Cure models are specifically designed for analysing such data by estimating cured proportions, understanding the lifetime distribution of uncured patients, and assessing the impact of covariates on lifetime (Maller and Zhou 1996).

There exist two classifications for cure models: the traditional two-component mixture cure model and the relatively recent promotion time cure model. The mixture cure model, pioneered by Boag (1949), further refined by Berkson and Gage (1952) and later explored by Wang et al. (2020) and Felizzi et al. (2021) is a prevalent model for estimating the survival rate of untreated patients and cure rate of a treatment simultaneously. As a widely favoured substitute to the mixture cure model, Tsodikov et al. (1996) and Tsodikov (1998) offered and investigated the promotion time cure model,

$$S_{pop}(t) = \exp\left[-\beta F(t)\right],\tag{1.1}$$

where $\beta > 0$ and $F(\cdot)$ is a proper baseline distribution function. It is also known as bounded cumulative hazard model since cumulative hazard $-\log S_{pop}(t) = \beta F(t) < \infty$. The model can be best understood through the following seminal biological interpretation of (1.1), based on a biological mechanism observed in cancer patients after treatment. The number of active carcinogenic cells remaining after treatment be denoted by N and modelled by a Poisson distribution having mean β . C_i denotes the time taken by the i^{th} cell to develop an identifiable cancerous growth, with T as their minimum. Assuming C_i s are independently distributed with common cumulative distribution function F(t), $S_{pop}(t)$ is the survival function of T, signifying the probability of being cancer-free by time t.

Chen et al. (1999) modified (1.1) for modelling the population survival function of an individual pertaining to covariate $\mathbf{X} = (1, X_1, \dots, X_k)'$, as

$$S_{pop}(t|\mathbf{X}) = P(T > t|\mathbf{X}) = \exp\left[-e^{\theta'\mathbf{X}}F(t)\right],$$
 (1.2)

by letting $\beta = e^{\theta' \mathbf{X}}$, where $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_k)'$ represents the vector of regression parameters. The immune or cured proportion according to model (1.2) equals

$$p(\mathbf{X}) = \lim_{t \to \infty} S_{pop}(t|\mathbf{X}) = \exp\left[-e^{\theta'\mathbf{X}}\right] > 0.$$
 (1.3)

It may be noted that, under the model (1.2), the covariates influence both the survival of patients who remain uncured and the probability of being cured. From a Bayesian perspective, the promotion time cure model outperforms the mixture cure model by yielding proper posterior distributions even with non-informative improper priors. Additionally, it features a proportional hazards structure and is biologically motivated, making it better suited for describing cancer relapse processes (Legrand and Bertrand 2019). Within the framework of the promotion time cure model, a nonparametric method for modelling the effects of covariates is introduced by Chen and Du (2018) and a semiparametric version of the model that accounts for nonlinear effects of a continuous covariate is explored by Lin and Huang (2024). Further, the efficiency and identifiability of the model have been examined and investigated by Portier et al. (2017) and Lambert and Bremhorst (2019) respectively. For recent research on promotion time cure model, one could refer to Gómez et al. (2023), Lin and Huang (2024), and Pal and Aselisewine (2023) among many others.

Some notable Bayesian inference procedures employing the promotion time cure model include those by Rahimzadeh and Kavehie (2016) and Gressani and Lambert (2018).

Limited literature exists on modelling current status data arising out of a population having an immune subgroup. Notable frequentist contributions include mixture cure models by Lam and Xue (2005) and Liu et al. (2017), a generalized linear model by Ma (2009), an additive risk model by Ma (2011), a non-mixture cure model by Wang and Han (2020), and transformation cure models by Diao and Yuan (2019) and Lam et al. (2021). The Bayesian approach utilising Markov Chain Monte Carlo (MCMC) methods, offers advantages over frequentist approaches by incorporating prior information and providing exact inference without relying on asymptotic approximations. To date, the only Bayesian cure models developed for interval-censored data are those by Ahmed (2021) and Pan et al. (2023). However, a Bayesian promotion time cure model for modelling current status data with covariates is not yet developed. Inspired by this research gap, we have devised a Bayesian method intended to analyse current status data, relying on a promotion time cure model.

The paper adheres to the following structure. Section 2 sets forth the Bayesian inference procedure, followed by techniques for model validation and comparison. Efficiency of the suggested method is inquired into by means of simulation studies and utility is exemplified with real datasets in Section 3 and 4 respectively. To conclude, certain observations are presented in Section 5.

2. Bayesian Inference Procedure

Consider the promotion time cure model (1.2). Our objective is to estimate the distribution of uncured population through F(t), covariate effects through $\boldsymbol{\theta}$ and the cured proportion $p(\mathbf{X})$ given by (1.3). Under current status censoring, assume non-informative censoring through the notion that, given \mathbf{X} , T and U are independent. Further suppose that the monitoring time distribution depends on none of the parameters of interest. Presuming n subjects in the study who are independent, we obtain the data, \mathcal{D} = $\{(U_i, \delta_i, \mathbf{X}_i) : i = 1, \ldots, n\}$. With \mathcal{D} , the associated likelihood is established as

$$L(\boldsymbol{\theta}, F(\cdot)|\mathcal{D}) = \prod_{i=1}^{n} S_{pop} \left(U_{i} | \mathbf{X}_{i} \right)^{1-\delta_{i}} \left(1 - S_{pop} \left(U_{i} | \mathbf{X}_{i} \right) \right)^{\delta_{i}}.$$
 (2.1)

Given assumption (1.2), (2.1) transforms into

$$L(\boldsymbol{\theta}, F(\cdot)|\mathcal{D}) = \prod_{i=1}^{n} \exp\left[-e^{\boldsymbol{\theta}'\mathbf{X}_{i}}F(U_{i})(1-\delta_{i})\right] \left(1 - \exp\left[-e^{\boldsymbol{\theta}'\mathbf{X}_{i}}F(U_{i})\right]\right)^{\delta_{i}}.$$
 (2.2)

2.1. Prior Distributions

Bayesian analysis builds on prior knowledge about the parameters in a statistical model, represented by the prior distributions, even before considering any data.

One may observe from (2.2) that merely the values of the continuous distribution function $F(\cdot)$ at U_i 's affect the likelihood function. Therefore, without loss of generality, one can focus only on the estimation of $F(\cdot)$ within the class of all non-decreasing step functions having discontinuities at distinct monitoring times, say $0 = s_0 < s_1 < s_2 < \cdots < s_{n_0}$. As presented by Sun (2006), consider the step functions of the form,

$$F_s(t) = 1 - \prod_{l:s_l < t} \exp(-e^{\eta_l}),$$
 (2.3)

with $\eta = (\eta_1, \dots, \eta_{n_0})'$, where $\eta_l = \log(-\log(r_l))$ and $r_l = \frac{1 - F_s(s_l)}{1 - F_s(s_{l-1})}$, for $l = 1, \dots, n_0$. For prior of η , an n_0 -variate normal distribution with mean $\mu = (\mu_1, \dots, \mu_{n_0})'$ is chosen. As its variance covariance matrix, Σ_{η} , an $n_0 \times n_0$ matrix with non-zero non-diagonal elements is selected, to account for the dependence of η_l ; $l = 1, \dots, n_0$ with its adjacent components. Thus, $\eta \sim N_{n_0}(\mu, \Sigma_{\eta})$ with the probability density function:

$$\pi_{\eta}(\eta) = \frac{1}{(2\pi)^{n_0/2}} |\Sigma_{\eta}|^{-1/2} e^{\frac{-1}{2}(\eta-\mu)'\Sigma_{\eta}^{-1}(\eta-\mu)},$$

where μ and Σ_{η} can be either known or estimated by the experimenter based on their expertise.

Assuming the independence of regression coefficients with real support, a (k+1)-variate normal distribution is adopted as the prior for $\boldsymbol{\theta}$. This distribution has mean vector $\boldsymbol{\tau} = (\tau_0, \tau_1, \dots, \tau_k)'$ and a $(k+1) \times (k+1)$ variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, having zeroes as non-diagonal elements, signifying independence between the coefficients. Therefore, $\boldsymbol{\theta} \sim N_{k+1}(\boldsymbol{\tau}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ characterised by its probability density function

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{(k+1)/2}} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-1/2} e^{\frac{-1}{2}(\boldsymbol{\theta} - \boldsymbol{\tau})' \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\tau})},$$

where τ and Σ_{θ} are chosen by the experimenter to incorporate prior knowledge or uncertainty regarding the regression parameters.

Additionally, the independence between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ is assumed, given the covariate vector \mathbf{X} .

Remark 2.1 The priors are selected based on experimental knowledge, with options ranging from non-informative priors that offer inferential priority to data, to informative

priors when prior knowledge is available.

2.2. Posterior Computation

The likelihood is coupled with the priors to form the posterior distribution, encapsulating improved beliefs about the parameters after observing the data.

Employing (2.3), the likelihood function (2.2) is rewritten in terms of θ and η by

$$L(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathcal{D}) = \prod_{i=1}^{n} \exp \left[-e^{\boldsymbol{\theta}' \mathbf{X}_{i}} \left(1 - \prod_{l: s_{l} \leq U_{i}} \exp(-e^{\eta_{l}}) \right) (1 - \delta_{i}) \right]$$

$$\times \left(1 - \exp \left[-e^{\boldsymbol{\theta}' \mathbf{X}_{i}} \left(1 - \prod_{l: s_{l} \leq U_{i}} \exp(-e^{\eta_{l}}) \right) \right] \right)^{\delta_{i}}.$$

$$(2.4)$$

Taking into account the dataset \mathcal{D} and postulates specified in Subsection 2.1, concerning the prior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ as well as their independence, posterior distribution $\pi^*(\cdot)$ is expressed as

$$\pi^*(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathcal{D}) \propto \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \pi_{\boldsymbol{\eta}}(\boldsymbol{\eta}) L(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathcal{D}).$$
 (2.5)

Adopting the squared error loss function, marginal posterior means of the parameters provide the Bayes estimators as

$$\tilde{\theta}_{\nu} = E_{\pi_{\nu}^{*}}(\theta_{\nu}|\mathcal{D}) = \int_{\theta_{0}} \int_{\theta_{1}} \cdots \int_{\theta_{\nu}} \cdots \int_{\theta_{k}} \theta_{\nu} \pi_{\theta}^{*}(\boldsymbol{\theta}|\mathcal{D}) d\theta_{0} d\theta_{1} \dots d\theta_{\nu} \dots d\theta_{k}; \nu = 0, 1, \dots, k,$$
(2.6)

and

$$\tilde{\eta}_{l} = E_{\pi_{l}^{*}}(\eta_{l}|\mathfrak{D}) = \int_{\eta_{1}} \int_{\eta_{2}} \cdots \int_{\eta_{l}} \cdots \int_{\eta_{n_{0}}} \eta_{l} \pi_{\boldsymbol{\eta}}^{*}(\boldsymbol{\eta}|\mathfrak{D}) d\eta_{1} d\eta_{2} \dots d\eta_{l} \dots d\eta_{n_{0}}; l = 1, \dots, n_{0},$$

$$(2.7)$$

where $\pi_{\boldsymbol{\theta}}^*(\boldsymbol{\theta}|\mathcal{D})$ and $\pi_{\boldsymbol{\eta}}^*(\boldsymbol{\eta}|\mathcal{D})$ denote the marginal posterior densities of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ respectively.

An approach for estimating F(t) utilising the Bayes estimators $\tilde{\eta}_l$, $l = 1, ..., n_0$, in (2.7), is proposed through the formulation;

$$\tilde{F}_s(t) = 1 - \prod_{l:s_l \le t} \exp\left(-e^{\tilde{\eta}_l}\right). \tag{2.8}$$

The estimator for the population survival function of a subject given the vector of co-

variates X is suggested as

$$\tilde{S}_{pop}(t|\mathbf{X}) = \exp\left(-e^{\tilde{\boldsymbol{\theta}}'\mathbf{X}} \left[1 - \prod_{l:s_l \le t} \exp\left(-e^{\tilde{\eta}_l}\right)\right]\right). \tag{2.9}$$

With the vector of covariates **X** and Bayes estimators from (2.6), $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_0, \tilde{\theta}_1, \dots, \tilde{\theta}_k)'$ in hand, the proposed estimator for the cure fraction is expressed as

$$\tilde{p}(\mathbf{X}) = \lim_{t \to \infty} \tilde{S}_{pop}(t|\mathbf{X}) = \exp\left(-e^{\tilde{\boldsymbol{\theta}}'\mathbf{X}}\right). \tag{2.10}$$

Theoretical assessment of (2.6) and (2.7) is complicated, prompting the utilisation of MCMC methods.

2.3. Posterior Simulation

The conditional posterior densities of parameters θ_{ν} ; $\nu=0,1,\ldots,k$ and η_l ; $l=1,\ldots,n_0$ do not possess closed forms and Gibbs sampling ceases to be applicable. Therefore, an adaptive Metropolis-Hastings (MH) algorithm (Haario et al. 1999) is deployed for posterior simulation utilising *MHadaptive* package in *R*-software with a few modifications. This sophisticated form of MH algorithm uses Gaussian proposal distribution with dynamic variance-covariance structure determined by process history. For more details and variations of the algorithm, see Chauveau and Vandekerkhove (2002), Cai et al. (2008), Griffin and Walker (2013), and Marnissi et al. (2020).

The algorithm outlined in the subsequent steps generates a Markov chain $(\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)})$ having $\pi^*(\cdot)$ as the approximate stationary distribution, where $\boldsymbol{\theta}^{(m)} = (\theta_0^{(m)}, \theta_1^{(m)}, \dots, \theta_k^{(m)})$ and $\boldsymbol{\eta}^{(m)} = (\eta_1^{(m)}, \dots, \eta_{n_0}^{(m)})$.

- (i) Formulate function (2.5) using the priors for θ , η , and the dataset \mathcal{D} .
- (ii) Set initial parameter values $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\eta}^{(0)})$ and calculate the parameter values $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\eta}^{(1)})$ that maximise (2.5), then set m = 1.
- (iii) Select the Gaussian proposal distribution, with the inverse of the observed Fisher information matrix evaluated at $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\eta}^{(1)})$ as the variance-covariance matrix and $(\boldsymbol{\theta}^{(m)'}, \boldsymbol{\eta}^{(m)'})'$ as mean.
- (iv) Generate new values of parameters $\boldsymbol{\theta}^{(m)p}$ and $\boldsymbol{\eta}^{(m)p}$ out of the proposal distribution considered. Then choose ω randomly from U(0,1).
- (v) Calculate the transition probability $\phi((\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)}), (\boldsymbol{\theta}^{(m)p}, \boldsymbol{\eta}^{(m)p}))$ as the minimum of $\left\{1, \frac{\pi^*(\boldsymbol{\theta}^{(m)p}, \boldsymbol{\eta}^{(m)p}|\mathcal{D})}{\pi^*(\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)}|\mathcal{D})}\right\}$. If $\log \ \omega$ is smaller than or equal to $\log \ \phi((\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)}), (\boldsymbol{\theta}^{(m)p}, \boldsymbol{\eta}^{(m)p}))$, update $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)p}$ and $\boldsymbol{\eta}^{(m+1)} = \boldsymbol{\eta}^{(m)p}$. Otherwise, set $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\eta}^{(m+1)} = \boldsymbol{\eta}^{(m)p}$.

- (vi) Increase m by one and execute steps (iii)-(v) for a predefined number of iterations, determined based on the Markov chain diagnostics. At specified intervals, the variance-covariance matrix of proposal distribution is updated adaptively using a portion of the previously generated values (Haario et al. 1999).
- (vii) The values of parameters, followed by an adequate burn-in and suitable thinning, produce the posterior sample $(\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)})$; $m = 1, \ldots, m_0$, that constitutes a nearly independent sample from a stationary distribution approximating $\pi^*(\cdot)$.

(viii) Calculate (2.6) and (2.7) as

$$\tilde{\theta}_{\nu} = \frac{1}{m_0} \sum_{m=1}^{m_0} \theta_{\nu}^{(m)}, \nu = 0, 1, \dots, k,$$
(2.11)

where $\theta_{\nu}^{(m)}$ represents the ν^{th} element of the vector $\boldsymbol{\theta}^{(m)}$ for $m=1,\ldots,m_0$.

$$\tilde{\eta}_l = \frac{1}{m_0} \sum_{m=1}^{m_0} \eta_l^{(m)}, l = 1, \dots, n_0,$$
(2.12)

where $\eta_l^{(m)}$ represents the l^{th} element of the vector $\boldsymbol{\eta}^{(m)}$ for $m=1,\ldots,m_0$.

The Ergodic theorem (Robert et al. 1999) ensures that the empirical averages (2.11) and (2.12) converge to integrals (2.6) and (2.7) respectively.

2.4. Model Comparison and Validation

Two common measures chosen for comparing multiple models from a Bayesian standpoint are the Log pseudo-marginal likelihood (LPML) and the Deviance Information Criterion (DIC) due to Geisser and Eddy (1979) and Spiegelhalter et al. (2002) respectively. Various researchers have devised expressions for these, across different semiparametric regression models, utilising current status data (Wang and Dunson 2011; Hariharan et al. 2023; Hariharan and Sankaran 2024).

LPML is viewed as an indicator of the overall predictive performance of the data, with higher values suggesting better performance. It is computed using Bayesian cross validated residual or conditional predictive ordinate values (CPO). CPO_i ; $i=1,\ldots,n$ represents the predictive probability of i^{th} observation taking into account leftover data; $\mathcal{D}^{(-i)} = \{(U_j, \delta_j, \mathbf{X}_j) : j = 1, \ldots, n, j \neq i\}$ assuming that the current model is valid. CPO_i is determied using

$$CPO_{i} = P(T_{i} \in (L_{i}, R_{i}] | \mathcal{D}^{(-i)})$$

$$= \left(E_{\pi^{*}} \left[\frac{1}{P(T_{i} \in (L_{i}, R_{i}] | \boldsymbol{\theta}, \boldsymbol{\eta})}\right]\right)^{-1},$$

 CPO_i due to Chen et al. (2012) upon obtaining the MCMC sample $(\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)}); m = 1, \ldots, m_0$ is

$$CPO_{i} = \left(\frac{1}{m_{0}} \sum_{m=1}^{m_{0}} \left[\frac{1}{P(T_{i} \in (L_{i}, R_{i}] | \boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)})}\right]\right)^{-1}$$

$$= \left(\frac{1}{m_{0}} \sum_{m=1}^{m_{0}} \left[\frac{1}{\delta_{i} + (-1)^{\delta_{i}} \exp\left(-e^{\boldsymbol{\theta}^{(\tilde{m})'} \mathbf{X}_{i}} \left[1 - \prod_{l: s_{l} \leq U_{i}} \exp\left(-e^{\tilde{\eta}^{(m)}}\right)\right]\right)\right]\right)^{-1}.$$
(2.13)

Subsequently, LPML is estimated as

$$LPML = \sum_{i=1}^{n} \log CPO_i.$$
 (2.14)

DIC serves as a technique for comparing Bayesian models. It strikes a balance between model fit and complexity, with lower values suggesting a better fit while minimising complexity. The DIC is computed as the sum of the expected value of deviance $D(\theta, \eta)$ and a penalty term p_D for complexity, as

$$DIC = E_{\pi^*}[D(\boldsymbol{\theta}, \boldsymbol{\eta})] + p_D,$$

where, $E_{\pi^*}(\cdot)$ refers to the expected value with regard to $\pi^*(\cdot)$ specified in (2.5), $D(\boldsymbol{\theta}, \boldsymbol{\eta}) = -2 \log L(\boldsymbol{\theta}, \boldsymbol{\eta}|data) + c$, with c a constant, and $p_D = E_{\pi^*}[D(\boldsymbol{\theta}, \boldsymbol{\eta})] - D[E_{\pi^*}(\boldsymbol{\theta}, \boldsymbol{\eta})]$. Therefore,

$$DIC = 2E_{\pi^*}[D(\boldsymbol{\theta}, \boldsymbol{\eta})] - D[E_{\pi^*}(\boldsymbol{\theta}, \boldsymbol{\eta})],$$

which can be approximated using sample $(\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)}); m = 1, \dots, m_0$, as

$$DIC = 2\tilde{D}(\boldsymbol{\theta}, \boldsymbol{\eta}) - D(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}}). \tag{2.15}$$

Here, $\tilde{D}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is the average deviance over m_0 samples and $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_0, \tilde{\theta}_1, ..., \tilde{\theta}_k)'$, $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, ..., \tilde{\eta}_{n_0})$ are obtained according to equations (2.11) and (2.12) respectively.

CPO also serves as a valuable tool for assessing model adequacy. By plotting CPO values or scaled CPO values (normalised to the maximum CPO value) against observation indices or monitoring times U_i ; i = 1, ..., n, analysts can evaluate model fit. Larger CPO values suggest better agreement between the model and observations, while low values (scaled CPO < 0.01) imply outliers and poor model fit to the data (Congdon 2005). A model is believed satisfactory if its CPOs or scaled CPOs exhibit a random distribution without outliers, as corroborated by Aslanidou et al. (1998) and Sinha and Maiti (2004).

3. Simulation Studies

Evaluation of the proposed Bayesian estimation procedure in estimating model parameters is undertaken through simulation studies. Two covariates, X_1 distributed as a Bernoulli distribution having probability of success 0.5 and X_2 following a standard normal distribution are presumed to impact the event times. Five distinct parameter combinations representing both positive and negative covariate effects are assumed, and 500 datasets of identical size are generated for each combination. The Bayesian estimation procedure outlined is then applied to each dataset, yielding parameter estimates. Estimations are performed with sample sizes (n) of 200 and 500 and posterior summaries are provided subsequently. Consider a simple model

$$S_{pop}(t|\mathbf{X}) = \exp\left(-e^{\theta_0 + \theta_1 X_1 + \theta_2 X_2} \left(1 - e^{-\frac{b}{a}(e^{at} - 1)}\right)\right), \tag{3.1}$$

where a, b > 0, $\theta_0, \theta_1, \theta_2$ are real numbers and $\mathbf{X} = (1, X_1, X_2)'$. A Gompertz distribution is used to model F(t) with survivor function $S(t) = e^{-\frac{b}{a}(e^{at}-1)}$ and (3.1) incorporates a proportion of immune subjects.

Implementing the method discussed in Oulhaj and Martin (2014) for generating data from any improper distribution, current status data with a cured fraction is produced using the model (3.1). To generate the required data, presume a = 0.5, b = 1.1 and keep $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$ fixed. For $i = 1, \ldots, n$, generate $\mathbf{X}_i = (1, X_{1i}, X_{2i})'$, where $X_{1i} \sim B(1, 0.5)$ and $X_{2i} \sim N(0, 1)$. Compute $1 - p(\mathbf{X}_i) = 1 - \exp(-e^{\boldsymbol{\theta}'\mathbf{X}_i})$ and generate a random number λ_i from $B(1, 1 - p(\mathbf{X}_i))$. If $\lambda_i = 0$, set $T_i = \infty$; if $\lambda_i = 1$, generate χ_i from Uniform(0, 1) and compute the event time T_i using,

$$T_i = S^{-1} \left[1 + \frac{1}{\exp(\boldsymbol{\theta}' \mathbf{X}_i)} \log \left[1 - \chi_i \left(1 - \exp\left(- \exp(\boldsymbol{\theta}' \mathbf{X}_i) \right) \right) \right] \right].$$

In scenario (1), a fixed censoring scheme is employed, selecting ten equidistant monitoring times $(s_1, \ldots, s_{10}) = (0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3.0)$, for ease of implementation. A multinomial random vector (n_1, \ldots, n_{10}) with all probabilities equal is generated, subject to $\sum_{l=1}^{10} n_l = n$. Decide the value of U_i using the pattern: set U_i equal to s_1 when i is between 1 and n_1 , equal to s_2 when i is between $n_1 + 1$ and $n_1 + n_2$, and so on, until U_i equals s_{10} when i is between $n_1 + n_2 + \ldots + n_9 + 1$ and n. For $i = 1, \ldots, n$, δ_i is set to 1 if T_i is smaller than or equal to U_i and otherwise set to zero. This method produces the current status data from a population with a cured fraction, represented as $\{(U_i, \delta_i, \mathbf{X}_i); i = 1, \ldots, n\}$. Additionally, under scenario (2) involving a random censoring scheme, ten non-equidistant monitoring times are generated from Uniform(0, 3), which is a more common occurrence in practice, and the entire process is then repeated.

To give inferential priority to data, vague prior $N(1, 10^2)$ is chosen for θ_0 , θ_1 , and θ_2 . Using $\mu_l = \log \left[-\log \left(\frac{S(s_l)}{S(s_{l-1})} \right) \right]$ for $l = 1, \ldots, 10$, $\boldsymbol{\mu} = (-1.03, -0.88, -0.73, -0.58, -0.43, -0.28, -0.13, 0.02, 0.17, 0.32)'$ is obtained. Noting that $\eta_l; l = 1, \ldots, n_0$, by definition, has the highest dependence with adjacent components and the lowest with distant ones, an appropriate choice for Σ_{η} is a first-order autoregressive structure $\Sigma_{n_0}(\rho)$, given by

$$\boldsymbol{\Sigma}_{n_0}(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_0-1} \\ \rho & 1 & \rho & \dots & \rho^{n_0-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n_0-3} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \rho^{n_0-1} & \rho^{n_0-2} & \rho^{n_0-3} & \dots & 1 \end{pmatrix},$$

where the correlation between adjacent elements is ρ , with $0 < \rho < 1$. Therefore, η is assumed to follow $N_{10}(\mu, \Sigma_{10}(0.3))$. The adaptive MH algorithm in Subsection 2.3 is applied with each of the 500 simulated datasets, yielding posterior estimates $\tilde{\theta}_{\nu}$; $\nu = 0, 1, 2$ and $\tilde{\eta}_l$; l = 1, ..., 10. The results are derived from 70,000 MCMC samples, following a burn-in period of 10,000 samples and thinning to every fifteenth to mitigate autocorrelation. Convergence diagnostics of MCMC are detailed in Appendix A.1. Approximately 0.015 seconds are required for each MCMC iteration, with n = 500.

Table 1 The results for regression parameters under scenario (1) and (2) with n = 200

	True	$\begin{bmatrix} (1) & (u_1, u_2,, u_{10}) = (0.3, 0.6,, 3.0) \end{bmatrix}$			(2) $u_i \sim Uniform(0,3)$						
	Truc	Mean	Abs. bias	EPSD	SSD	CP	Mean	Abs. bias	EPSD	SSD	CP
θ_0	0.6	0.6189	0.0189	0.1736	0.1814	0.93	0.5618	0.0382	0.1737	0.1730	0.95
θ_1	-0.5	-0.5107	0.0107	0.2119	0.2192	0.95	-0.5167	0.0167	0.2155	0.2154	0.97
θ_2	0.7	0.7474	0.0474	0.1314	0.1452	0.94	0.7398	0.0398	0.1320	0.1468	0.92
θ_0	-0.8	-0.8534	0.0534	0.2231	0.2445	0.93	-0.8894	0.0894	0.2297	0.2625	0.93
θ_1	-1	-1.0725	0.0725	0.3186	0.3304	0.96	-1.0261	0.0261	0.3224	0.3240	0.96
θ_2	-1.2	-1.2565	0.0565	0.1967	0.2084	0.96	-1.2482	0.0482	0.1973	0.2132	0.94
θ_0	-0.75	-0.7929	0.0429	0.2269	0.2264	0.94	-0.8310	0.0810	0.2317	0.2361	0.94
θ_1	2.1	2.2154	0.1154	0.3548	0.3955	0.94	2.1810	0.0810	0.3508	0.3698	0.94
θ_2	1.5	1.6091	0.1091	0.2398	0.2744	0.95	1.5796	0.0796	0.2350	0.2688	0.93
θ_0	-1.5	-1.6270	0.1270	0.2964	0.3072	0.96	-1.6594	0.1594	0.3034	0.3738	0.92
θ_1	1.7	1.8543	0.1543	0.3526	0.3862	0.94	1.7910	0.0910	0.3512	0.3883	0.92
θ_2	-1.9	-2.0636	0.1636	0.2861	0.3221	0.94	-2.0172	0.1172	0.2820	0.3396	0.92
θ_0	-1	-1.0782	0.0782	0.2543	0.2545	0.96	-1.0918	0.0918	0.2604	0.2993	0.92
θ_1	-1.25	-1.3345	0.0845	0.3592	0.4029	0.92	-1.3303	0.0803	0.3609	0.3762	0.97
θ_2	1.75	1.8775	0.1275	0.2717	0.3251	0.92	1.8453	0.0953	0.2688	0.3094	0.94

Tables 1 and 2 present the frequentist operating characteristics of $\tilde{\theta}_0$, $\tilde{\theta}_1$, and $\tilde{\theta}_2$ under scenarios (1) $(s_1, \ldots, s_{10}) = (0.3, \ldots, 3)$ and (2) $s_i \sim U(0, 3)$, for n = 200 and n = 500

Table 2 The results for regression parameters under scenario (1) and (2) with n = 500

	True	$ (1) (u_1, u_2,, u_{10}) = (0.3, 0.6,, 3.0) $			(2) $u_i \sim Uniform(0,3)$						
	True	Mean	Abs. bias	EPSD	SSD	CP	Mean	Abs. bias	EPSD	SSD	CP
θ_0	0.6	0.5971	0.0029	0.1104	0.0975	0.98	0.5374	0.0266	0.1132	0.1111	0.95
θ_1	-0.5	-0.5064	0.0064	0.1314	0.1382	0.94	-0.5103	0.0103	0.1335	0.1387	0.94
θ_2	0.7	0.7161	0.0161	0.0804	0.0860	0.93	0.6978	0.0022	0.0803	0.0783	0.96
θ_0	-0.8	-0.8056	0.0056	0.1440	0.1403	0.96	-0.8628	0.0628	0.1501	0.1660	0.93
θ_1	-1	-1.0423	0.0423	0.1948	0.2198	0.92	-1.0165	0.0165	0.1999	0.1980	0.94
θ_2	-1.2	-1.2288	0.0288	0.1192	0.1211	0.96	-1.2354	0.0354	0.1227	0.1223	0.96
θ_0	-0.75	-0.7820	0.0320	0.1479	0.1476	0.97	-0.7926	0.0426	0.1494	0.1735	0.93
θ_1	2.1	2.1455	0.0455	0.2156	0.2295	0.96	2.0954	0.0046	0.2154	0.2274	0.94
θ_2	1.5	1.5416	0.0416	0.1447	0.1714	0.92	1.5244	0.0244	0.1439	0.1537	0.94
θ_0	-1.5	-1.5561	0.0561	0.1851	0.1989	0.94	-1.5589	0.0589	0.1869	0.2024	0.92
θ_1	1.7	1.7603	0.0603	0.2137	0.2317	0.94	1.7262	0.0262	0.2141	0.2124	0.94
θ_2	-1.9	-1.9790	0.0790	0.1725	0.2045	0.92	-1.9332	0.0332	0.1699	0.1925	0.94
θ_0	-1	-1.0200	0.0200	0.1632	0.1517	0.93	-1.0416	0.0416	0.1679	0.1703	0.96
θ_1	-1.25	-1.2962	0.0462	0.2177	0.2386	0.93	-1.2726	0.0226	0.2228	0.2339	0.94
θ_2	1.75	1.7898	0.0398	0.1616	0.1612	0.95	1.7668	0.0168	0.1621	0.1756	0.94

respectively. Here, 'Mean' is the average of 500 posterior mean estimates. Absolute value of bias is denoted by Abs. bias. The standard deviation of every posterior sample is calculated and these estimated posterior standard deviations are averaged across all replications to generate EPSD. Additionally, SSD is determined by computing the sample standard deviation of the posterior mean estimates. Percentile approach is opted to form 95% Bayesian credible intervals (BCI) and the percentage of these BCIs encompassing the true value is given by CP, the coverage probability. The efficacy of the proposed method under both fixed and random censoring schemes for both the sample sizes is evident as the mean values closely approximate the true values of θ_0 , θ_1 , and θ_2 , EPSD and SSD exhibit minimal differences as well as low magnitudes, and CP values are close to 0.95. With increase in sample size, reduction of Abs. bias, EPSD and SSD is also observed.

Table 3 The maximum value among the local MSEs of $\tilde{F}_s(t)$ under scenario (1) and (2) with n=200

	(0.6, -0.5, 0.7)	(-0.8,-1,-1.2)	(-0.75,2.1,1.5)	(-1.5,1.7,-1.9)	(-1,-1.25,1.75)
(1)	0.1712	0.1713	0.2072	0.1814	0.1961
(2)	0.2391	0.3281	0.2596	0.3012	0.2181

Using estimates of $\eta_l; l=1,\ldots,10$, obtained from five different settings, $\tilde{F}_s(t)$ are computed by (2.8) at these ten distinct monitoring times. The effectiveness of the proposed method in estimating distribution function is evaluated by computing the local mean square errors of $\tilde{F}_s(t)$ at 10 distinct monitoring times. Tables 3 and 4 summarise

Table 4 The maximum value among the local MSEs of $\tilde{F}_s(t)$ under scenario (1) and (2) with n = 500

	(0.6, -0.5, 0.7)	(-0.8,-1,-1.2)	(-0.75, 2.1, 1.5)	(-1.5,1.7,-1.9)	(-1,-1.25,1.75)
(1)	0.0906	0.1399	0.1574	0.1178	0.1419
(2)	0.1203	0.1939	0.1301	0.1868	0.2073

the maximum of these errors (MaxMSE) for n = 200 and n = 500 respectively, which are consistently minimal across all setups and smaller for larger samples, indicating accurate estimation of F(t). Additionally, Fig. 1 illustrates the comparison between the actual F(t) curves and the estimated curves (constructed by connecting the estimates at ten specific monitoring times, allowing for a direct comparison against the true F(t) curves) for three setups, demonstrating the consistent performance of the proposed estimation procedure.

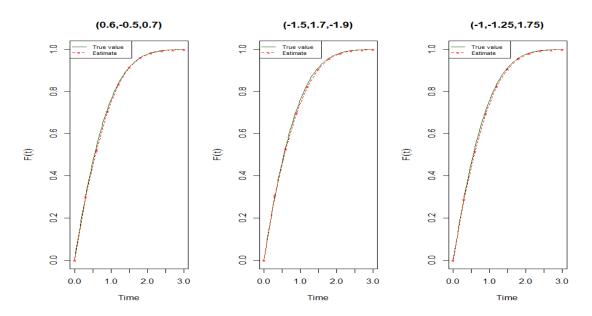


Fig. 1 The estimates of F(t) alongside the true curve

4. Illustrative Examples

4.1. Lung Tumor Data

Sun (2006) has presented the data from a tumorigenicity experiment conducted by Hoel (1972) involving 144 RFM mice. The study is designed to evaluate the influence of environmental conditions on the time to lung tumor onset (T_i) in mice. As tumor onset is not directly observable, only the time of death or sacrifice (U_i) and an indicator of the presence or absence of a tumor at that time (δ_i) are recorded, resulting in current status data. Since lung tumors in RFM mice are non-lethal, tumor onset has not affected their mortality and the event time T_i remains independent of the examination time U_i .

The experiment involved placing 96 mice in a conventional environment (CE) and 48 in a germ-free environment (GE). The median follow-up time is 662.5 days and the event times are censored for 82 mice (71.9% in CE, 27.1% in GE). To assess the presence of a cured fraction, the non-parametric maximum likelihood estimator (NPMLE) of the survival function, $\hat{S}(\cdot)$, was plotted for both environments, as shown in Fig. 2. A long plateau at the tail of $\hat{S}(\cdot)$ for CE indicates the presence of a cure fraction among the mice in the conventional environment. According to Maller and Zhou (1996), one may look for fitting a cure model to the data in order to estimate the impact of environment on the time to tumor onset and the cure fraction.

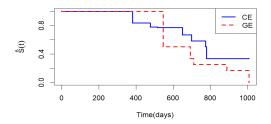


Fig. 2 NPMLE of S(t) for lung tumor data

In order to begin with the analysis, the data is modified by selecting eleven distinct monitoring times $(s_1, \ldots, s_{11}) = (45, 381, 477, 515, 650, 679, 773, 779, 839, 888, 1008)$ each representing the time intervals during which $\hat{S}(\cdot)$ remains piecewise constant. As an informative prior for η , $N_{11}(\mu, 0.1 * \Sigma_{11}(0.3))$ is chosen, where $\mu_l = \log\left(-\log\left(\frac{\hat{S}(s_l)}{\hat{S}(s_{l-1})}\right)\right)$ for $l = 1, \ldots, 11$. Wang and Han (2020) has introduced a binary covariate X to represent the environment, assigning 0 for CE and 1 for GE, and has applied maximum likelihood estimation to fit equation (1.2) to the data. Building on this, informative normal priors centered at their MLEs with minimum spread are used for the regression coefficients: $\theta_0 \sim N(-0.27, 0.01^2)$ and $\theta_1 \sim N(0.81, 0.01^2)$. Posterior computation is performed with the adaptive MH algorithm (see Subsection 2.3) and the results are reported in Table 5, with MCMC diagnostics detailed in Appendix A.2.

Table 5 Summary of Bayesian estimates for the lung tumor data

Parameters	Estimates	Posterior standard deviations	BCI
$ heta_0$	-0.2702	0.0100	(-0.2906,-0.2502)
$ heta_1$	0.8102	0.0102	(0.7899, 0.8298)

The estimates for θ_0 (negative) and θ_1 (positive) are statistically significant, as their BCIs exclude zero. This indicates that the environment significantly affects lung tumor risk, with mice in the germ-free environment (GE) being at higher risk. Using (2.10), the cure rates are estimated as 0.4662 for CE and 0.1798 for GE, indicating that mice in CE are more likely to be cured. The population survival curves, estimated using (2.9)

and shown in Fig. 3 (constructed by connecting the estimates at distinct observational times), further confirm that survival probabilities are higher for mice in CE compared to GE. Fig. 4 shows scaled CPOs plotted against the mice indices, with no visible pattern

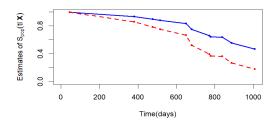


Fig. 3 Estimates of population survival function for lung tumor data

or outliers, suggesting a good model fit. The LPML and DIC of the model are -87.96 and 91.08 respectively.

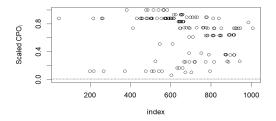


Fig. 4 Scaled CPO plot for lung tumor data

4.2. Breast Cancer Data

Lösch et al. (1998) gathered data from 100 women with primary invasive ductal carcinoma, who underwent initial surgical treatment, followed by monitoring for the first confirmed metastasis or recurrence. This breast cancer dataset utilised by Heinze and Schemper (2001) and now accessible in the coxphf package of R software, provides information on the recurrence-free interval, representing the duration between the initial surgical procedure and the first verified metastasis or recurrence, subject to right censoring. Additionally, it incorporates four potential risk factors: X_1 =tumor stage (i.e. 1, if stage is 2, 3 or 4 and 0, if stage is 1), X_2 =histological grading (i.e. 1, if grade is 2 or 3 and 0, if grade is 1), X_3 =nodal status (i.e. 1, if number of nodes is 1 or 2 and 0, if number of nodes is 0), and X_4 =cathepsin D (CD) immunoreactivity (i.e. 1, if CD positive and 0, if CD negative), intended for evaluating their impact on survival time.

The median follow-up time for the patients is 72 months. Among the 100 patients, 74% are censored, with 51% of the censoring occurring at the maximum follow-up time

of 72 months. According to Maller and Zhou (1996), The presence of several censored observations near the largest monitoring time and the Kaplan-Meier estimator's plateau around 0.714 in Fig. 5 indicate the potential presence of a cure fraction and the suitability of a cure model. To illustrate the proposed Bayesian estimation procedure, the data are transformed into current status data, by grouping lifetime data into 12 non-overlapping intervals ([0,6), [6,12), [12,18), [18,24), [24,30), [30,36), [36,42), [42,48), [48,54), [54,60), [60,66), and [66,72]) and converting these into exact observational times by considering their midpoints, simplifying the implementation.

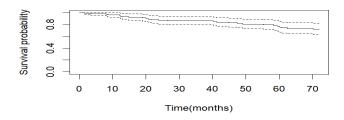


Fig. 5 Kaplan-Meier Survival curve of breast cancer data

Unlike the previous example, a promotion time cure model has not yet been fitted to the data to avail some prior information regarding the parameter values. Therefore, careful prior elicitation is crucial to reduce uncertainty around the parameters and ensure meaningful posterior probabilities. To determine the priors for regression coefficients, cure probabilities are obtained using the Kaplan-Meier curve of the survival function and a method for prior elicitation outlined in Lambert and Bremhorst (2019) is implemented. This resulted in the informative priors, $\theta_0 \sim N(-1.171, 0.212^2)$, $\theta_1 \sim N(0.737, 0.044^2)$, $\theta_2 \sim N(0.378, 0.002^2)$, $\theta_3 \sim N(0.789, 0.077^2)$, and $\theta_4 \sim N(0.587, 0.095^2)$. In order to obtain a prior for η , the Kaplan-Meier estimator of the baseline survival function at distinct monitoring times s_l ; $l=1,\ldots,12$ are noted as $\hat{S}_{pop}(s_l|\mathbf{X}=0)$ and an estimator for F(.) at s_l is obtained using $\hat{F}(s_l) = \frac{-\log \hat{S}_{pop}(t|\mathbf{X}=0)}{mean(\theta_0)}$ for $l=1,\ldots,12$. Further, $N_{12}(\boldsymbol{\mu},\boldsymbol{\Sigma}_{12}(0.3))$ is adopted as $\boldsymbol{\eta}$ -prior, with $\mu_l = \log\left(-\log\left(\frac{1-\hat{F}(s_l)}{1-\hat{F}(s_{l-1})}\right)\right)$ for $l=1,\ldots,12$. Posterior summary of the parameter estimation on implementing the proposed adaptive MH algorithm is shown in Table 6. One can see Appendix A.3 for more details on the computation as well as Markov chain diagnostics.

The estimates of all regression coefficients except intercept are positive and statistically significant, as none of their BCIs include zero. Using these estimates, the cure rates for patients at various levels of covariates X_j ; j = 1, 2, 3, 4 can be estimated using (2.10). For instance, an individual diagnosed with tumor stage 1 and being CD positive has a cure rate of 0.5559, which is lower than the baseline cure rate of 0.8474, estimated by setting all the covariates at level zero. Additionally, one can observe that the cure rates at

Table 6 Summary of Bayesian estimates with the breast cancer data

Parameters	Estimates	Posterior standard deviations	BCI
θ_0	-1.7986	0.1366	(-2.0721, -1.5332)
θ_1	0.7311	0.0459	(0.6402, 0.8202)
θ_2	0.3778	0.0025	(0.3729, 0.3829)
θ_3	0.7718	0.0753	(0.6315, 0.9250)
θ_4	0.5553	0.0896	(0.3756, 0.7252)

unfavorable levels of covariates are lower than the baseline cure rate. These observations affirm that the patients in higher stage of tumor, higher levels of grading, having higher number of nodes, and CD positivity are less likely to cure the risk of tumor recurrence in comparison with others.

In addition to the cure rate, the population survival functions are estimated as shown in Fig. 6, demonstrating that unfavorable levels of covariates are associated with lower survival probabilities. Therefore, all four risk factors are recognised as prognostic markers in the context of breast cancer, with higher tumor stage and grading, increased number of involved lymph nodes, and positivity for Cathepsin D pertaining to a significantly negative prognostic impact on survival. This information enhances our comprehension of the biological attributes of breast cancer and aids in patient management and treatment decision-making.

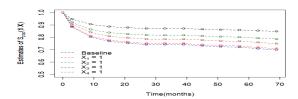


Fig. 6 Estimates of population survival function for breast cancer data

Furthermore, the plot of the Bayesian estimates of the baseline survival function in Fig. 6 closely matches the Kaplan-Meier curve for actual data in Fig. 5. The random pattern of scaled CPOs in Fig. 7 suggests that the model fits to the data adequately. The model's LPML and DIC are -50.57 and 99.68 respectively.

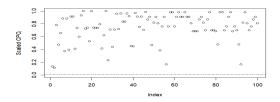


Fig. 7 Scaled CPO plot for breast cancer data

5. Conclusions

The paper has presented a comprehensive framework for a Bayesian promotion time cure model, designed to analyse current status data. The Bayesian estimation procedure adopted in this study is notable for its utilisation of proper priors for the model parameters and an adaptive Metropolis-Hastings algorithm for posterior computation. This model not only advances the existing methods of estimation, but also introduces novel techniques for model comparison and validation using cross-validated predictive ordinate (CPO). The introduced model contributes significantly to the Bayesian cure model literature, as established through simulation results and practical data analyses. By incorporating prior knowledge and uncertainty, this approach yields more reliable parameter estimates, thereby enhancing the overall robustness and applicability of cure modelling techniques.

In the literature, Box-Cox-type transformations including specific cases like the proportional hazards and proportional odds cure models are used in cure rate models to enhance model flexibility. Bayesian techniques can be devised for these models to offer robust parameter estimation and uncertainty quantification. The current status censoring becomes informative when time of examination is associated with the time of event. To address this issue and prevent potential bias, a shared frailty model can be employed, capturing the interplay between event occurrence time and time of examination. Our intention is to further investigate Bayesian extensions in this direction, aiming to refine our understanding of complex disease dynamics and improve model performance in clinical research.

Acknowledgements

The first author wishes to acknowledge the financial support of the Council of Scientific & Industrial Research, Government of India, via the Junior Research Fellowship scheme under reference No. 09/0239(13499)/2022-EMR-I.

References

Ahmed, A. O. M. (2021). Bayesian estimations of exponential distribution based on interval-censored data with a cure fraction. *Journal of Mathematics*, 2021(1):9822870.

Al-Mosawi, R. and Lu, X. (2022). Efficient estimation of semiparametric varying-coefficient partially linear transformation model with current status data. *Journal of Statistical Computation and Simulation*, 92(2):416–435.

- Andersen, P. K. and Ronn, B. B. (1995). A nonparametric test for comparing two samples where all observations are either left-or right-censored. *Biometrics*, 51(1):323–329.
- Aslanidou, H., Dey, D. K., and Sinha, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods. *Canadian Journal of Statistics*, 26(1):33–48.
- Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. Journal of the American Statistical Association, 47(259):501–515.
- Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53.
- Cai, B., Lin, X., and Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics & Data Analysis*, 55(9):2644–2651.
- Cai, B., Meyer, R., and Perron, F. (2008). Metropolis–Hastings algorithms with adaptive proposals. *Statistics and Computing*, 18:421–433.
- Chauveau, D. and Vandekerkhove, P. (2002). Improving convergence of the Hastings—Metropolis algorithm with an adaptive proposal. *Scandinavian Journal of Statistics*, 29(1):13–29.
- Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.
- Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo Methods in Bayesian Computation*. Springer Science & Business Media, New York.
- Chen, T. and Du, P. (2018). Promotion time cure rate model with nonparametric form of covariate effects. *Statistics in Medicine*, 37(10):1625–1635.
- Congdon, P. (2005). Bayesian Models for Categorical Data. John Wiley & Sons, New York.
- Das, S., Chae, M., Pati, D., and Bandyopadhyay, D. (2024). Bayesian semiparametric modeling of spatially-referenced multistate current status data. In *APHA 2024 Annual Meeting and Expo.* APHA.
- Diamond, I. D., McDonald, J. W., and Shah, I. H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography*, 23(4):607–620.

- Diao, G. and Yuan, A. (2019). A class of semiparametric cure models with current status data. *Lifetime Data Analysis*, 25(1):26–51.
- Felizzi, F., Paracha, N., Pöhlmann, J., and Ray, J. (2021). Mixture cure models in oncology: a tutorial and practical guidance. *Pharmacoecon Open*, 5(2):143–155.
- Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal* of the American Statistical Association, 74(365):153–160.
- Gómez, Y. M., Gallardo, D. I., Bourguignon, M., Bertolli, E., and Calsavara, V. F. (2023). A general class of promotion time cure rate models with a new biological interpretation. Lifetime Data Analysis, 29(1):66–86.
- Gressani, O. and Lambert, P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics & Data Analysis*, 124:151–167.
- Griffin, J. E. and Walker, S. G. (2013). On adaptive Metropolis–Hastings methods. Statistics and Computing, 23:123–134.
- Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Springer Science & Business Media, New York.
- Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–395.
- Hariharan, P. and Sankaran, P. G. (2024). Semiparametric regression modelling of current status competing risks data: a Bayesian approach. *Computational Statistics*, 39:2083–2108.
- Hariharan, P., Sankaran, P. G., and Mulayath Variyath, A. (2023). A Bayesian semi-parametric regression model for current status data. *Communications in Statistics-Simulation and Computation*, https://doi.org/10.1080/03610918.2023.2266153.
- Heinze, G. and Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, 57(1):114–119.
- Heung, M., Wolfgram, D. F., Kommareddi, M., Hu, Y., Song, P. X., and Ojo, A. O. (2012). Fluid overload at initiation of renal replacement therapy is associated with lack of renal recovery in patients with acute kidney injury. *Nephrology Dialysis Transplantation*, 27(3):956–961.
- Hoel, D. G. (1972). A representation of mortality data by competing risks. *Biometrics*, 28(2):475–488.

- Jewell, N. P. and Shiboski, S. C. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics*, 46(4):1133–1150.
- Jewell, N. P. and van der Laan, M. (2003). Current status data: Review, recent developments and open problems. *Handbook of Statistics*, 23:625–642.
- Lam, K. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika*, 92(3):573–586.
- Lam, K. F., Lee, C. Y., Wong, K. Y., and Bandyopadhyay, D. (2021). Marginal analysis of current status data with informative cluster size using a class of semiparametric transformation cure models. *Statistics in Medicine*, 40(10):2400–2412.
- Lambert, P. and Bremhorst, V. (2019). Estimation and identification issues in the promotion time cure model when the same covariates influence long-and short-term survival. Biometrical Journal, 61(2):275–289.
- Laughlin, S. K., Baird, D. D., Savitz, D. A., Herring, A. H., and Hartmann, K. E. (2009). Prevalence of uterine leiomyomas in the first trimester of pregnancy: an ultrasound screening study. *Obstet Gynecol*, 113(3):630–635.
- Legrand, C. and Bertrand, A. (2019). Cure models in cancer clinical trials. In *Textbook* of clinical trials in oncology, pages 465–492. Chapman and Hall/CRC.
- Li, S., Hu, T., Wang, L., McMahan, C. S., and Tebbs, J. M. (2024). Regression analysis of group-tested current status data. *Biometrika*, 111(3):1047–1061.
- Li, S., Tian, T., Hu, T., and Sun, J. (2021). A simulation-extrapolation approach for regression analysis of misclassified current status data with the additive hazards model. *Statistics in Medicine*, 40(28):6309–6320.
- Lin, L.-H. and Huang, L.-S. (2024). Promotion time cure model with local polynomial estimation. Statistics in Biosciences, https://doi.org/10.1007/s12561-024-09423-y.
- Liu, Y., Hu, T., and Sun, J. (2017). Regression analysis of current status data in the presence of a cured subgroup and dependent censoring. *Lifetime Data Analysis*, 23:626–650.
- Lösch, A., Tempfer, C., Kohlberger, P., Joura, E., Denk, M., Zajic, B., Breitenecker, G., and Kainz, C. (1998). Prognostic value of cathepsin D expression and association with histomorphological subtypes in breast cancer. *British Journal of Cancer*, 78(2):205–209.
- Ma, S. (2009). Cure model with current status data. Statistica Sinica, 19(1):233–249.

- Ma, S. (2011). Additive risk model for current status data with a cured subgroup. *Annals of the Institute of Statistical Mathematics*, 63(1):117–134.
- Maller, R. A. and Zhou, X. (1996). Survival Analysis with Long-Term Survivors. John Wiley & Sons, New York.
- Marnissi, Y., Chouzenoux, E., Benazza-Benyahia, A., and Pesquet, J.-C. (2020). Majorize-minimize adapted Metropolis-Hastings algorithm. *IEEE Transactions on Signal Processing*, 68:2356–2369.
- Oulhaj, A. and Martin, E. S. (2014). Generating data from improper distributions: application to Cox proportional hazards models with cure. *Journal of Statistical Computation and Simulation*, 84(1):204–214.
- Pal, S. and Aselisewine, W. (2023). A semiparametric promotion time cure model with support vector machine. *The Annals of Applied Statistics*, 17(3):2680–2699.
- Pan, C., Cai, B., and Sui, X. (2023). A Bayesian proportional hazards mixture cure model for interval-censored data. *Lifetime Data Analysis*, https://doi.org/10.1007/s10985-023-09613-8.
- Paulon, G., Müller, P., and Sal y Rosas, V. G. (2024). Bayesian nonparametric bivariate survival regression for current status data. *Bayesian Analysis*, 19(1):49–75.
- Portier, F., El Ghouch, A., and Van Keilegom, I. (2017). Efficiency and bootstrap in the promotion time cure model. *Bernoulli*, 23(4B):3437–3468.
- Rahimzadeh, M. and Kavehie, B. (2016). Promotion time cure model with generalized Poisson-inverse Gaussian distribution. *Journal of Biostatistics and Epidemiology*, 2(2):68–75.
- Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.
- Sinha, D. and Maiti, T. (2004). A Bayesian approach for the analysis of panel-count data with dependent termination. *Biometrics*, 60(1):34–40.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.
- Sun, J. (2006). The Statistical Analysis of Interval-Censored Failure Time Data. Springer, New York.

- Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. Biometrics, 54(4):1508–1516.
- Tsodikov, A. D., Yakovlev, A. Y., and Asselain, B. (1996). Stochastic Models of Tumor Latency and Their Biostatistical Applications. World Scientific, Singapore.
- Wang, C. and Du, M. (2024). Martingale-residual-based greedy model averaging for high-dimensional current status data. *Statistics in Medicine*, 43(9):1726–1742.
- Wang, L. and Dunson, D. B. (2011). Semiparametric Bayes' proportional odds models for current status data with underreporting. *Biometrics*, 67(3):1111–1118.
- Wang, X. and Han, B. (2020). Efficient estimation for the non-mixture cure model with current status data. *Statistics*, 54(4):756–777.
- Wang, Y., Tang, Y., and Zhang, J. (2020). Bayesian approach for proportional hazards mixture cure model allowing non-curable competing risk. *Journal of Statistical Computation and Simulation*, 90(4):638–656.
- Wu, Q., Tong, X., and Zhao, X. (2024). Deep partially linear cox model for current status data. *Biometrics*, https://doi.org/10.1093/biomtc/ujae024, 80(2).

Conflicts of Interest

There are no conflicts of interest between the authors.

Appendix A. MCMC Convergence and Mixing Diagnostics

Convergence of Markov chains for simulation studies and real data analyses is demonstrated within the Appendix. As graphical checks, autocorrelation plots (ACF plots), trace plots, and posterior histograms are examined. Gelman-Rubin diagnostics, effective sample sizes (ESS), and acceptance rate are also reported.

Appendix A.1. Simulation Studies

Consider $(\theta_0, \theta_1, \theta_2) = (0.6, -0.5, 0.7)$ under scenario (1). With a randomly generated data, 70,000 MCMC simulations are done. As burn-in, 10,000 samples are removed and the remaining ones are thinned, keeping only multiples of 15. ACF plots in Fig. 8 visualise chain autocorrelation: high autocorrelation implies poor mixing, while low

values indicate better convergence. The plotted parameters show rapid autocorrelation decay, indicating well-behaved simulated Markov chains. Trace plots in Fig. 9 depict the evolution of MCMC-generated samples for parameters of interest. The plots show random fluctuations around a central value without anomalies, suggesting well-mixed posterior samples. Posterior histograms in Fig. 10 are stable with consistent shapes and narrow peaks. This indicates convergence and low uncertainty in parameter estimation. Gelman-Rubin diagnostics assesses convergence across multiple chains. Values of potential scale reduction factor close to 1 for θ_0 , θ_1 , and θ_2 indicate convergence, as observed with ten independent chains. The ESS for θ_0 , θ_1 , and θ_2 are 618, 616, and 673 respectively. Every MCMC repetition requires approximately 0.015 seconds. The rate of acceptance is 0.0969.

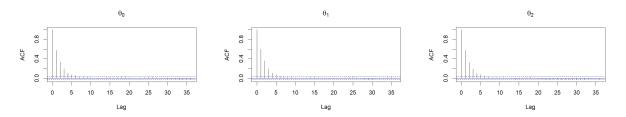


Fig. 8 ACF plots of parameters when $(\theta_0, \theta_1, \theta_2) = (0.6, -0.5, 0.7)$

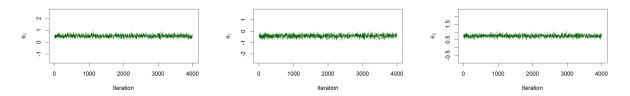


Fig. 9 Trace plots of parameters when $(\theta_0, \theta_1, \theta_2) = (0.6, -0.5, 0.7)$

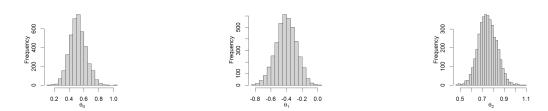


Fig. 10 Posterior histograms of parameters when $(\theta_0, \theta_1, \theta_2) = (0.6, -0.5, 0.7)$

Appendix A.2. Lung Tumor Data Analysis

In the analysis of lung tumor data, Markov chain diagnostics employ 50,000 MCMC samples, with a burn-in of 10,000 and retention of every 15th sample. Various graphical assessments are illustrated in Fig. 11, 12, and 13. The Gelman-Rubin diagnostic values,

approaching 1, indicate that the chains have converged effectively. The ESS for θ_0 and θ_1 are 399 and 500 respectively. Each iteration takes 0.0108 seconds to compute, with an acceptance rate of 0.1072.

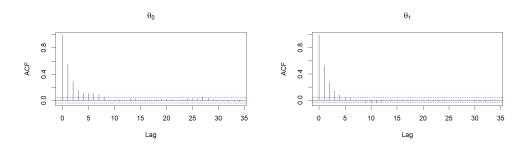


Fig. 11 ACF plots: Lung tumor data analysis

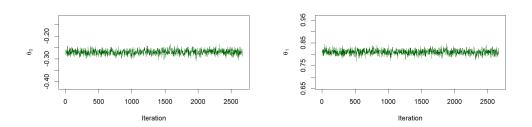


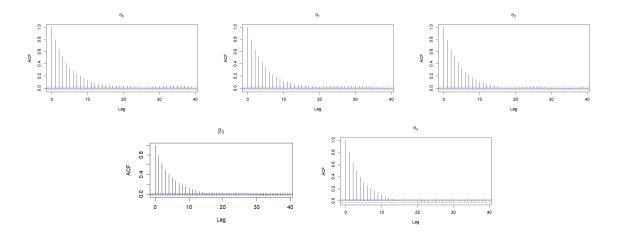
Fig. 12 Trace plots: Lung tumor data analysis



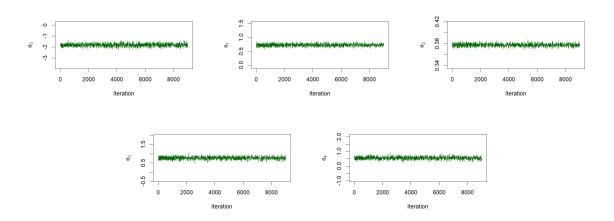
Fig. 13 Posterior histograms: Lung tumor data analysis

Appendix A.3. Breast Cancer Data Analysis

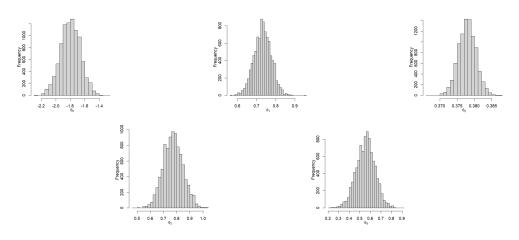
In the breast cancer data analysis, Markov chain diagnostics utilise 100,000 MCMC samples, with 10,000 as burn-in and retaining only the multiples of 10. Fig. 14, 15, and 16 demonstrate various graphical checks. Gelman-Rubin diagnostics values close to 1 further confirm the chains' convergence. ESS for θ_0 , θ_1 , θ_2 , θ_3 , and θ_4 are 708, 732, 805, 740, and 873 respectively. Computing time is 0.0152 seconds per iteration and acceptance rate is 0.0602.



 ${\bf Fig.~14}$ ACF plots: Breast cancer data analysis



 ${\bf Fig.}~{\bf 15}$ Trace plots: Breast cancer data analysis



 ${\bf Fig.~16}$ Posterior histograms: Breast cancer data analysis