

HHS Public Access

Stat Methods Med Res. Author manuscript; available in PMC 2021 November 01.

Published in final edited form as:

Author manuscript

Stat Methods Med Res. 2020 November; 29(11): 3192-3204. doi:10.1177/0962280220921552.

A Bayesian approach for analyzing partly interval-censored data under the proportional hazards model

Chun Pan¹, Bo Cai², Lianming Wang³

¹Department of Mathematics and Statistics, Hunter Collge, New York, NY 10065

²Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29208

³Department of Statistics, University of South Carolina, Columbia, SC 29208

Abstract

Partly interval-censored time-to-event data often occur in biomedical studies of diseases where periodic medical examinations for symptoms of interest are necessary. Recent decades have seen blooming methods and R packages for interval-censored data, however, the research effort for partly interval-censored data is limited. We propose an efficient and easy-to-implement Bayesian semiparametric method for analyzing partly interval-censored data under the proportional hazards model. Two simulation studies are conducted to compare the performance of the proposed method with two main Bayesian methods currently available in the literature and the classic Cox proportional hazards model. The proposed method is applied to a partly interval-censored progression-free survival data from a metastatic colorectal cancer trial.

Keywords

Bayesian semiparametric; partly interval-censored; proportional hazards model; progression-free survival

1 Introduction

Partly interval-censored data often occur in medical and health studies that include periodic examinations. With partly interval-censored data, the failure times are exactly observed for some subjects, while only known to be within certain time intervals for the rest. In cancer clinical trials, progression-free survival, defined as time from study entry to disease progression or death due to any cause, is often used as the primary endpoint. It is actually partly interval-censored as the exact date of death is normally known while the date of disease progression is only known to be between two assessment visits. The mainstream methods in pharmaceutical industry are to ignore this so-called arbitrary censoring attribute of the data and continue to treat it as right-censored by assuming that the event occurs at the

Corresponding author: Chun Pan, 695 Park Avenue, New York, NY 10065, cp2677@hunter.cuny.edu, Tel.: 212-396-6040. Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

study day when it is detected. This strategy can induce bias in the estimation especially when the intervals are wide and varied.^{1,2} The standard error of the estimation is also underestimated since it assumes that failure times are exactly known when they are not.³

The current literature for partly interval-censored data is limited. From the frequentist perspective, Huang⁴ developed the asymptotic properties for the nonparametric maximum likelihood estimator (NPMLE) of the distribution function in Turnbull's⁵ model. Kim³ developed the maximum likelihood estimator for the proportional hazards (PH) model. Zhao et al.⁶ developed a class of generalized log-rank tests that perform survival comparison. Gao et al.² developed semiparametric estimation of the accelerated failure time (AFT) model. From the Bayesian perspective, Zhou and Hanson⁷ developed a unified approach that fits PH, proportional odds, and AFT models to partly interval-censored and left-truncated spatial data. The two functions that implement their method for partly interval-censored data are survregbayes and survregbayes2 in their R package **spBayesSurv**.⁸ Komárek and Lesaffre⁹ developed a mixed-effects AFT model for partly interval-censored data. This method is implemented by the bayessurvreg1 function in their R package **bayesSurv**.¹⁰

It is worth mentioning that there are R packages for interval-censored data that can also be used for fitting partly interval-censored data. They include the **intcox** package that implements Pan's method¹¹ which extends the iterative convex minorant algorithm to the Cox PH model for interval-censored data; the **MIICD** package¹² that implements multiple imputation for PH regression with interval-censored data; the **coarseDataTools** package¹³ that fits parametric AFT models to interval-censored data; the **interval** package¹⁴ that estimates the NPMLE of survival curve and performs log-rank and Wilcoxon type tests for interval-censored data; the **SmoothHazard** package¹⁵ that can fit semiparametric or parametric PH model to interval-censored data; the **survBayes** package¹⁶ that fits a PH model by a Bayesian approach to interval-censored data; and the **icenReg** package¹⁸ that fits Bayesian PH models of and AFT models for interval-censored data.

However, there might be limitations with some of these packages. For instance, the **intcox** package does not provide standard error for an estimated regression coefficient. The **interval** package does not perform regression analysis. The **MIICD** package imputes exact times for finite interval-censored data and and then use the partial likelihood method. The **survBayes** package reduces the data to right-censored data by imputing an observed time for each finite interval-censored time. The **dynsurv** package also reduces the data to "augmented right-censored data" through sampling exact times for finite interval-censored times. The **coarseDataTools** package uses the survreg function in the **survival** package¹⁹ or uses the general optimization function optim and reduces intervals to their midpoints. For Bayesian inference, the **icenReg** package fits parametric models only. Further evaluation of the performance of these packages under partly interval-censored data may also be helpful.

In this paper, we introduce an efficient and easy-to-implement Bayesian approach specifically developed for analyzing partly interval-censored data under the semiparametric PH model. The main differences between the proposed method and the two Bayesian methods we compare with are: (1) Zhou and Hanson used the transformed Bernstein

Page 3

polynomial prior or mixtures of Polya trees prior to model the baseline survival function; while we use a mixture of basis I-splines to model the baseline cumulative hazard function. (2) Zhou and Hanson used an adaptive Metropolis sampler²⁰ to sample regression coefficients; while we use the Metropolis-Hastings algorithm²¹ to sample regression coefficients. (3) Komárek and Lesaffre fit an AFT model with the error term specified as a normal mixture with an unknown number of components; while we fit the PH model. Zhou and Hanson has pointed out that models using the mixtures of Polya trees prior can suffer from poor mixing and the transformed Bernstein polynomial prior is preferred. According to the discussions by Diaconis and Ylvisaker²² and Perron and Mengersen,²³ the approximation based on Bernstein polynomials can be poor for some nonlinear functions. On the other hand, the adaptive Metropolis algorithm samples a vector of parameters with proposal variance $\frac{2.4^2}{d}C_t + 10^{-10}I_d$, where C_t is the sample variance of all previous draws, d is the dimension of the vector sampled, and I_d is the identity matrix. As a multidimensional sampler, it poses more difficulty in achieving convergence to target distribution and good mixing. Simulation II in Section 3 demonstrates one scenario where the proposed method outperforms Zhou and Hanson's method.

The remainder of the paper is outlined as follows. Section 2 describes the proposed method including spline approximation, data augmentation, prior specification, and posterior computation. Section 3 presents two simulation studies that evaluate the performance of the method and compares it with Zhou and Hanson and Komárek and Lesaffre for partly interval-censored data, and the classic Cox PH model.²⁴ In Section 4, we derive partly interval-censored progression-free survival data based on the overall tumor responses from a phase III metastatic colorectal cancer trial and compared the analysis result by the proposed method with those from the other methods. Finally Section 5 provides conclusions and discussions.

2 Statistical method

2.1 Data structure and notation

Partly interval-censored data consist of exact event times and general interval-censored event times. Note that general interval-censored data include left-censored, interval-censored, and right-censored observations. The corresponding observed time intervals are $(0, R_i]$, $(L_i, R_i]$, and $(L_i, \infty]$. The proposed method can accommodate any of exact, left-censored, right-censored, interval-censored times, and a mixture of them. Let n_1 be the number of observations that are observed exactly and n_2 the number of general interval-censored observations. We have a total of $N = n_1 + n_2$ observations. Without loss of generality, for the first n_1 subjects, the failure times T_i , $i = 1, ..., n_1$ are exactly known, but for the other n_2 subjects, the failure times are only known to be within a time interval, denoted as $(L_i, R_i]$, $i = n_1 + 1, ..., N$, where L_i can be 0 and R_i can be ∞ . So the observed data are $\{(T_i, X_i)\}_{i=1}^{n_1}$ and $\{(L_i, R_i, X_i)\}_{i=n_1+1}^N$, where X_i is the *i*th subject's covariate vector.

We assume that failure time T and examination times are independent given the covariate vector **X**.

2.2 Model

Let $\lambda_0(t_i)$ denote the unspecified baseline hazard function, $\boldsymbol{\beta}$ the $p \times 1$ vector of regression coefficients, \mathbf{x}_i the $p \times 1$ covariate vector. Under the Cox proportional hazards model, the hazard $\lambda(t_i | \mathbf{x}_i)$ of a failure time *T* is proportional to the baseline hazard:

$$\lambda(t_i \mid \mathbf{x}_i) = \lambda_0(t_i) \exp(\boldsymbol{\beta}' \mathbf{x}_i). \tag{1}$$

For an exact observation, T_i is observed, and its likelihood function is

$$L_{1i}\{\boldsymbol{\beta}, \lambda_0(\cdot)\} = f(t_i \mid \mathbf{x}_i) = \lambda_0(t_i)\exp(\boldsymbol{\beta}'\mathbf{x}_i)\exp\{-\Lambda_0(t_i)\exp(\boldsymbol{\beta}'\mathbf{x}_i)\},$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ is the cumulative baseline hazard function.

For a general interval-censored observation, $(L_i, R_i]$ is the observed time interval, and its likelihood function is

$$L_{2i}\{\boldsymbol{\beta}, \lambda_0(\cdot)\} = \{F(R_i \mid \mathbf{x}_i)\}^{\delta_{1i}}\{F(R_i \mid \mathbf{x}_i) - F(L_i \mid \mathbf{x}_i)\}^{\delta_{2i}}\{1 - F(L_i \mid \mathbf{x}_i)\}^{\delta_{3i}},$$
(2)

where $F(t \mid x) = 1 - \exp(-\int_0^t \lambda(s \mid x) ds)$ is the cumulative distribution function given **x** and δ_1 , δ_2 , δ_3 are the left-, interval-, and right-censoring indicators. So the overall likelihood function is:

$$L\{\boldsymbol{\beta}, \lambda_0(\cdot)\} = \prod_{i=1}^{n_1} L_{1i}\{\boldsymbol{\beta}, \lambda_0(\cdot)\} \prod_{i=n_1+1}^{N} L_{2i}\{\boldsymbol{\beta}, \lambda_0(\cdot)\}.$$
(3)

2.3 Estimation of $\Lambda_0(t)$ and $\lambda_0(t)$

Following Cai et al., Pan et al., Lin et al., and Pan et al.,^{25–28} we model the cumulative baseline hazard function $\Lambda_0(t)$ with a linear combination of a set of basis I-splines:³³

$$\Lambda_0(t) = \sum_{l=1}^{K} \gamma_l I_l(t) \tag{4}$$

where $\{\gamma_l\}$ is a set of non-negative coefficients and $\{I_l(t)\}$ is a set of basis I-splines.

To construct the set of basis I-splines, we need to specify the degree (1=linear, 2=quadratic, 3=cubic, etc.) of each basis I-spline and an increasing sequence of knots within the data range. The set of basis I-splines are fully determined once the degree and the knots are specified. The number of basis I-splines (*K*) equals the degree plus the number of interior knots. In general, we recommend taking 2 or 3 as the degree value for adequate smoothness and 10–30 equally spaced knots for adequate modeling flexibility.

Note that knots and degree can be adjusted based on data. Together with the coefficients for the basis I-splines, a monotone spline created this way can provide great flexibility for

approximating a curve. Furthermore, the shrinkage prior for the spline coefficients γ_I as shown in Section 2.5 serves to: (1) keep those important basis functions and leave those unnecessary ones out; (2) avoid over-fitting problems that may be caused by using too many knots for flexibility.

For the baseline hazard function $\lambda_0(t)$, we model it with a linear combination of a set of basis M-splines:³³

$$\lambda_0(t) = \sum_{l=1}^K \gamma_l M_l(t),$$

where $\{\gamma_l\}$ is the same set of non-negative coefficients as in (4) and $\{M_k(t)\}$ is a set of basis M-splines. I-splines are the integrated M-splines such that $\{I_l(t) = \int_0^t M_l(s)ds\}$. In our model, they share the same knots and an I-spline of degree *k* corresponds to an M-spline of degree *k* - 1.

2.4 Data augmentation

Although one may use the Metropolis-Hastings algorithm to sample all the parameters from their posteriors based on the original data likelihood (3), it is difficult to find good proposal distributions to obtain reasonable acceptance rates and well mixed Markov chain Monte Carlo (MCMC) chains. To facilitate posterior computation, we construct the following data augmentations.

For general interval-censored data, a two-step data augmentation is constructed by taking advantage of the PH model structure and the spline modeling form of $\Lambda_0(t)$ in (4). Assume that there is an underlying recurrent event E, for which the number of occurrences N(t) within time interval (0, t] is a nonhomogeneous Poisson process with cumulative intensity function $\Lambda_0(t) \exp(\mathbf{\beta}' \mathbf{x})$. Define $T = \inf\{t : N(t) > 0\}$, time of first occurrence in the Poisson process. Then we have $P(T > t) = P(N(t) = 0) = \exp\{-\Lambda_0(t) \exp(\mathbf{\beta}' \mathbf{x})\}$, which is our survival function of interest. So *T* indeed follows the PH model in (1).

Now define two time points t_1 and t_2 such that $0 < t_1 < t_2$. For left-censored observations (0, R], we set $t_1 = R$ and t_2 undefined as long as greater than t_1 . For interval-censored observations (L, R], we set $t_1 = L$ and $t_2 = R$. For right-censored observations (L, ∞), we set $t_2 = L$ and t_1 undefined as long as less than t_2 . It is clear that $N(t_1)$ denotes the number of occurrences of E until time t_1 , and $N(t_2) - N(t_1)$ denotes the number of occurrences of E during the interval (t_1, t_2]. By the properties of nonhomogeneous Poisson process, the random variable $Z = N(t_1) \sim \text{Poi}(\exp(\Lambda_0(t_1) \exp(\boldsymbol{\beta' x})))$, the random variable $W = N(t_2) - N(t_1) \sim \text{Poi}(\exp(\{\Lambda_0(t_2) - \Lambda_0(t_1)\} \exp(\boldsymbol{\beta' x}))))$, and they are independent. For left-censored data, since t_2 is some point greater than $t_1 = R$, W can take any value and will not contribute any information about the failure time T. For interval-censored data, Z = 0 and W > 0. For right-censored data, t_1 is some point less than $t_2 = L$, so Z = W = 0. The augmented data likelihood function for subject i is

$$L_{2aug1, i}(\theta \mid Z_i, W_i) = \text{Poi}(Z_i)\text{Poi}(W_i)^{\delta_2 i} + \delta_3 i$$

 $\times \{1(Z_i > 0)\}^{\delta_1 i} \{1(Z_i = 0)1(W_i > 0)\}^{\delta_2 i} \{1(Z_i = 0)1(W_i = 0)\}^{\delta_3 i},$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda_0(\cdot))$ denotes the set of parameters, $1(\cdot)$ the indicator function, and $0^0 = 1$. Integrating out Z_i and W_i will lead to the original likelihood function in (2).

Furthermore, based on the additive property of Poisson distribution and the linear combination form of (4), decompose Z and W respectively into K independent Poisson latent variables $\{Z_l\}$ and $\{W_l\}$, such that $Z = \sum_{l=1}^{K} Z_l$ with $Z_l \sim \text{Poi}(\gamma_l I_k(t_1) \exp(\boldsymbol{\beta' x}))$ and $W = \sum_{l=1}^{K} W_l$ with $W_l \sim \text{Poi}(\{\gamma_l I_k(t_2) - \gamma_l I_k(t_1)\} \exp(\boldsymbol{\beta' x}))$, with constraints $\sum_{l=1}^{K} Z_l > 0$ if $\delta_1 = 1$, $\sum_{l=1}^{K} Z_l = 0$ and $\sum_{l=1}^{K} W_l > 0$ if $\delta_2 = 1$, and $\sum_{l=1}^{K} Z_l = \sum_{l=1}^{K} W_l = 0$ if $\delta_3 = 1$. Then for subject *i*, the further augmented data likelihood function is

$$L_{2aug2, i}(\theta \mid Z'_{il}s, W'_{il}s) = \left\{ \prod_{l=1}^{K} \text{Poi}(Z_{il})\text{Poi}(W_{il})^{\delta_{2i} + \delta_{3i}} \right\}$$

 $\times \{1(Z_i > 0)\}^{\delta_{1i}} \{1(Z_i = 0)1(W_i > 0)\}^{\delta_{2i}} \{1(Z_i = 0)1(W_i = 0)\}^{\delta_{3i}}.$

The likelihood function is simply a product of Poisson probability mass functions, which leads to relatively straightforward posterior computation to be presented in Section 2.5.

For exact times, it will be challenging to sample the basis M-spline coefficients γ_I directly given the summation form in the likelihood function:

$$L_1(\boldsymbol{\theta}) = \prod_{i=1}^{n_1} \left[\left\{ \sum_{l=1}^{K} \gamma_l M_l(t_i) \right\} \exp(\boldsymbol{\beta}' \mathbf{x}_i) \exp\{-\Lambda_0(t_i) \exp(\boldsymbol{\beta}' \mathbf{x}_i) \right\} \right]$$

We introduce latent variables $\mathbf{u}_i = (u_{i1}, u_{i2}, ..., u_{iK})$ ~Multinomial $\left(1; \frac{1}{K}, \frac{1}{K}, ..., \frac{1}{K}\right)$, then we can derive the augmented data likelihood function for the part of exact observations as:

$$L_{1aug}(\boldsymbol{\theta} \mid \mathbf{u}_{i}'s) = \prod_{i=1}^{n_{1}} \left[\{ K \prod_{l=1}^{K} (\gamma_{l} M_{l}(t_{i}))^{\boldsymbol{u}_{i}l} \} \exp(\boldsymbol{\beta}' \mathbf{x}_{i}) \exp\{-\Lambda_{0}(t_{i}) \exp(\boldsymbol{\beta}' \mathbf{x}_{i}) \} \right]$$

Integrating out \mathbf{u}_i 's will lead to the original likelihood function $L_1(\boldsymbol{\theta})$. Under this format, we can obtain a Gamma posterior distribution for each γ_i , i = 1, ..., K.

2.5 Prior specification and posterior computation

For spline coefficients, we assign an Exponential prior $\text{Exp}(\eta)$ for γ_I and a Gamma hyperpiror $\text{Ga}(a_{\eta}, b_{\eta})$ for η . This specification leads to conjugate posteriors for both γ_I and η . For a numeric (continuous or count) covariate, we assign a Normal prior $N(0, \sigma_0^2)$ for β_r . Since the corresponding posterior is not conjugate, the Metropolis-Hastings algorithm is used for sampling from the posterior. For a categorical covariate with *c* levels, we represent

it using c - 1 dummy variables. The Metropolis-Hastings algorithm as a general sampler can be used for sampling for its β_r too. However, here we treat a categorical covariate differently. The reason is that by specifying a Gamma prior Ga (a_{ϕ}, b_{ϕ}) for $\phi_r = \exp(\beta_r)$, the resulting posterior happens to be Gamma which can be directly sampled from and renders better MCMC chains. Then we transform ϕ_r back to β_r .

After initializing values for the parameters, the proposed MCMC algorithm proceeds in the following steps.

1. Let $Z_i = 0$ and $W_i = 0$ for all i, $Z_{il} = 0$ and $W_{il} = 0$ for all i and l. If $\delta_{1i} = 1$, then sample

$$\begin{split} & Z_i \sim \text{Poi}(\Lambda_0(R_i) \text{exp}(\beta' \mathbf{x}_i)) 1(Z_i > 0), \\ & (Z_{i1}, ..., Z_{iK}) \sim \text{Multinomial}(Z_i; p_{i1}, ..., p_{iK}), \\ & \text{and} \ (p_{i1}, ..., p_{iK}) \propto (\gamma_1 I_1(R_i), ..., \gamma_K I_K(R_i)). \end{split}$$

If $\delta_{2i} = 1$, then sample

$$\begin{split} & W_i \sim \text{Poi}(\{\Lambda_0(R_i) - \Lambda_0(L_i)\} \exp(\beta' \mathbf{x}_i)) 1(W_i > 0), \\ & (W_{i1}, ..., W_{iK}) \sim \text{Multinomial}(W_i; q_{i1}, ..., q_{iK}), \\ & \text{and} \; (q_{i1}, ..., q_{iK}) \propto (\gamma_1 \{I_1(R_i) - I_1(L_i)\}, ..., \gamma_K \{I_K(R_i) - I_K(L_i)\}). \end{split}$$

- 2. Sample $(u_{i1}, ..., u_{iK}) \sim$ Multinomial $(1; o_{i1}, ..., o_{iK})$ and $(o_{i1}, ..., o_{iK}) \propto (\gamma_1 M_1(t_i), ..., \gamma_K M_K(t_i))$.
- 3. For β_r corresponding to a numeric covariate, use the Metropolis-Hastings algorithm to sample from its full conditional distribution

$$p(\beta_{r} \mid Z_{i}'s, W_{i}'s, \beta_{-r}) \propto \exp\left[\sum_{i=1}^{n_{1}} \left\{ x_{ir}\beta_{r} - \Lambda_{0}(t_{i})e^{\beta'\mathbf{x}_{i}} \right\} \right] \\ \times \exp\left[\sum_{i=n_{1}+1}^{N} \left\{ x_{ir}\beta_{r}(Z_{i}\delta_{1i} + W_{i}\delta_{2i}) - e^{\beta'\mathbf{x}_{i}}(\Lambda_{0}(R_{i})(\delta_{1i} + \delta_{2i}) + \Lambda_{0}(L_{i})\delta_{3i}) \right\} \right] p(\beta_{r}),$$

where $p(\beta_r) = N(0, \sigma_0^2)$ is the prior used for β_r , and β_{-r} denotes all the β 's except for β_r .

4. For β_r corresponding to a categorical covariate, let $\phi_r = \exp(\beta_r)$, sample ϕ_r from

$$\begin{aligned} &\operatorname{Ga}(a_{\phi} + \sum_{i=1}^{n_{1}} x_{ir} + \sum_{i=n_{1}+1}^{N} x_{ir}(Z_{i}\delta_{1i} + W_{i}\delta_{2i}), b_{\phi} + \sum_{i=1}^{n_{1}} \Lambda_{0}(t_{i})e^{\beta'_{-r}\mathbf{x}_{i}}, -r_{x_{ir}} \\ &+ \sum_{i=n_{1}+1}^{N} e^{\beta'_{-r}\mathbf{x}_{i}}, -r\{\Lambda_{0}(R_{i})(\delta_{1i} + \delta_{2i}) + \Lambda_{0}(L_{i})\delta_{3i}\}x_{ir}), \end{aligned}$$

where $\mathbf{x}_{i,-r}$ is the covariate vector except for x_{ir} for subject *i*.

5. Sample γ_l , l = 1, ..., K, from

$$Ga(1 + \sum_{i=1}^{n_1} u_{il} + \sum_{i=n_1+1}^{N} (Z_{il}\delta_{1i} + W_{il}\delta_{2i}), \eta + \sum_{i=m_1+1}^{N} e^{\beta' \mathbf{x}_i} \{ I_l(R_i)(\delta_{1i} + \delta_{2i}) + I_l(L_i)\delta_{3i} \}$$

6. Sample η from Ga $(a_{\eta} + K, b_{\eta} + \sum_{l=1}^{K} \gamma_{l})$.

As we can see, latent variables and spline coefficients all can be sampled from standard distributions. Special sampling method (here Metropolis-Hastings) is only required for the regression coefficient of a numeric covariate.

3 Simulations

3.1 Simulation I

We evaluate the performance of the proposed method through a simulation study. A total of 100 data sets were generated. For each data set, the failure times were generated from the following PH model:

$$S(t \mid x_1, x_2) = \exp\{-\Lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2)\},\$$

where $\Lambda_0(t) = \log(1 + t)$, $\beta_1 = 1$, $\beta_2 = 1$, x_1 's ~ N(0, 0.5²), and x_2 's ~ Bernoulli(0.5). Note that x_1 and x_2 were independently sampled and a new set of covariates were generated for each data set. We assume that the random number of medical examinations performed for each person is 1 plus a Poisson random number with mean 2. The gap times between adjacent medical examinations follow an Exponential distribution with mean 1. The observed interval is formed by the consecutive examination times (including 0 and ∞) that contain the true failure time. In each data set, there are N = 460 subjects, around 20% of which are set to have exact event times observed.

To construct the basis I-splines and basis M-splines, we set the degree as 2 for the basis I-splines and chose 15 equally spaced knots within the range of observed times. For hyperparameters, we tried $\sigma_0^2 = 10$, 10, 100, 1000, $a_\eta = b_\eta = 0.01$, 0.1, 1, and $a_\phi = b_\phi = 0.01$, 0.1, 1. The results were very similar and we chose to use $\sigma_0^2 = 100$, $a_\eta = b_\eta = 1$, and $a_\phi = b_\phi = 1$. Fast convergence and good mixing were observed for all key parameters. For each MCMC chain, we set total number of iterations = 11,000, burn-in = 1000, and thin = 1.

We fit the proposed method, and compare it survregbayes and survregbayes2 in the **spBayesSurv** package, and bayessurvreg1 in the **bayesSurv** package. We also treat finite interval-censored data as exact data by taking the right endpoints as the event times, as has conventionally done by practitioners, and then fit the Cox PH model using the coxph function in the **survival** package. The purpose is to demonstrate the potential bias this conventional approach might introduce.

Table 1 summarizes the simulation results. For each parameter, the point estimate is the average of the 100 posterior means, the sample standard deviation (SSD) is the sample standard deviation of the 100 posterior means, the empirical standard error (ESE) is the average of the 100 estimated standard errors, and the 95% coverage probability (95CP) is the percentage of the 100 credible intervals for each β_r that contains the true parameter value. Effective sample size (ESS) and absolute value of Geweke's Z-score were computed based on the MCMC chains using the **coda** package.³⁷ Negative log-likelihood (NLLK) is the negative of log psuedo marginal likelihood from survregbayes and survregbayes2 and the negative of log-likelihood from coxph. Log-likelihood at each iteration from bayessurvreg1 was averaged to calculate negative log-likelihood. Deviance information criterion (DIC) was not calculated for bayessurvreg1 because the error variance at each iteration was not available.

As seen in Table 1, the proposed method, survregbayes, and survregbayes2 all perform very well. The survregbayes and survregbayes2 functions have relatively high effective sample size, however, the proposed method shows lower absolute Geweke's Z-score, deviance information criterion, and negative log-likelihood which indicate better MCMC convergence to the stationary distribution and better model goodness-of-fit. Note that bayessurvreg1 fits a Bayesian AFT model (log(T_i) = $\beta' \mathbf{x}_i + \epsilon_i$), so it makes sense that the estimated regression coefficients are of negative signs and the coverage probabilities are not presented. The estimation from coxph shows large bias, low coverage probability, and large negative log-likelihood, which illustrates the bias it can induce if we treat partly interval-censored data as right-censored data.

We also estimated baseline survival function $S_0(t)$ based on the 100 simulated data sets. The estimated baseline survival functions and the true baseline survival function are plotted in Figure 1. All four partly interval-censored methods provide good approximations to the true baseline survival. However, the estimated curve based on coxph deviates significantly from the true curve.

3.2 Simulation II

To explore more scenarios, we performed another simulation study where the true cumulative baseline hazard function is set to be $\Lambda_0(t) = t^2$. Compared to Simulation I where $\Lambda_0(t) = \log(1 + t)$, the risk of failure is much higher under the new function. Other settings are exactly the same as in Simulation I.

Table 2 summarizes the simulation results. As we can see, the proposed method provides the best estimation with small biases and coverage probabilities close to the nominal level. The effective sample size from the proposed method is low compared to the other three partly interval-censored methods, indicating relatively high autocorrelation among the MCMC samples. The point estimates and coverage probabilities from survregbayes and survregbayes2 are not very good, even though they have high effective sample sizes. The estimation from coxph deviates even further from the true values compared to Simulation I, which indicates that the conventional method might lead to more bias for analyzing partly interval-censored data from diseases that have fast failure rate.

We also plotted the estimated baseline survival function versus the true in Figure 2. The proposed method provides the best approximation to the true baseline survival, followed by bayessurvreg1, survregbayes, and survregbayes2. The estimated curve from coxph is still noticeably different from the true curve.

4 An application to progression-free survival data

We apply the proposed method to a randomized phase III study that compares the efficacy of FOLFIRI versus panitumumab + FOLFIRI in patients with previously treated metastatic colorectal cancer. FOLFIRI is a combination of chemotherapy drugs: fluorouracil, leucovorin, and irinotecan. Panitumumab is a fully human monoclonal antibody specific to the epidermal growth factor receptor. The primary endpoint is progression-free survival. Two binary covariates are of interest: treatment arm (FOLFIRI vs. panitumumab + FOLFIRI) and patient tumor KRAS mutation status (wild-type vs. mutant). KRAS stands for the gene Kirsten rat sarcoma viral oncogene homolog. It is one of a group of genes involved in the epidermal growth factor receptor pathway.

Both treatments were administered every 2 weeks. The visit schedule for tumor response evaluation was every 8 weeks until documentation of disease progression. Based on the overall tumor responses (e.g., complete response, partial response, stable disease, and progressive disease) at each visit across their on-study period, we derived the progression-free survival for each patient. We set baseline assessment as Day 0. If a patient had disease progression at the first post-baseline assessment, then he is left-censored. If a patient had disease progression at a later assessment, then he is interval-censored. If a patient was alive without disease progression at the last on-study assessment, then he is right-censored. If a patient died while on-study, then his progression-free survival is exact. After excluding 30 test failures and 61 missing values for KRAS mutation status, the final data set contains N= 855 patients, among which 52 died on-study, 168 left-censored, 329 interval-censored, and 306 right-censored. The FOLFIRI arm contains 427 randomly assigned patients with 234 wild-type and 193 mutant, while the panitumumab + FOLFIRI arm contains 428 randomly assigned patients with 240 wild-type and 188 mutant.

The estimation results using the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph are presented in Table 3. The proposed method, survregbayes, bayessurvreg1, and coxph all detect improvement of progression-free survival by adding panitumumab. The survregbayes2 function fails to detect a significant treatment effect. The four partly interval-censored methods have much lower negative log-likelihood than coxph, which indicates they fit the data better by taking into account that disease progression occurs between two visits instead of on a visit day.

Figure 3 presents the estimated survival curves for the four groups formed by treatment arm and KRAS mutation status based on the five methods compared in Table 2 and the classic Kaplan-Meier method.³⁹ As indicated by the estimated regression coefficients, the survival expectation is the highest for wild-type patients receiving panitumumab + FOLFIRI, followed by mutant patients receiving panitumumab + FOLFIRI, wild-type patients receiving FOLFIRI, and finally mutant patients receiving FOLFIRI.

Since panitumumab is an antibody targeted at the epidermal growth factor receptor and KRAS mutation status predicts the efficacy of such type of agents in metastatic colorectal cancer,^{40,41} we also compared the efficacy of panitumumab + FOLFIRI vs. FOLFIRI among patients with wild-type KRAS tumors as well as among patients with mutant KRAS tumors. Peeters et al.^{40,41} treated progression-free survival as right-censored and used the classic log-rank test and Cox PH model, but stratified by performance status, prior bevacizumab, and prior oxaliplatin exposure.

The results using the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph are summarized in Table 4. For wild-type patients, when patitumumab was added to FOLFIRI, a significant improvement in progression-free survival was observed based on all of the five methods. The results are consistent with that from Peeters et al.^{40,41} For mutant patients, only the proposed method detects a weak improvement in efficacy. The 95% CI from Peeters et al.⁴⁰ is (-0.386, 0.058), which indicates a non-significant trend toward increased progression-free survival. As in the first set of analysis, the proposed method has much higher effective sample size, indicating better mixing and more efficiency in generating effective samples.

5 Conclusion

In the past few decades, many statistical methods and R packages have been developed for interval-censored data. There have been limited research specifically developed for partly interval-censored data which also occur often in medical studies. The several methods developed from the frequentist perspective seem to be hard to implement by practitioners, or at least with no ready-to-use code available. The main Bayesian methods are the two R packages we have compared the proposed method to in this article: one fits PH, proportional odds, and AFT models to partly interval-censored data and left-truncated data and the other fits mixed effects AFT model to partly interval-censored data. We developed an efficient and easy-to-implement Bayesian semiparametric method under the PH model directly targeted at analyzing partly interval-censored data. The proposed method performs comparably well in terms of regression coefficient estimation and survival function estimation. It even outperforms the two R packages when the rate of failure is high as seen in Simulation II. Our developed method is a meaningful addition to the literature and we hope to provide pharmaceutical companies with another ready-to-use tool for analyzing partly intervalcensored data that are commonly encountered in cancer clinical trials, e.g. progression-free survival and disease-free survival.

Our simulation and real data analysis show that, when there is only one covariate, the effective sample size of the proposed method is pretty high. However, it may have less ideal mixing when there are more than one covariate. This is largely due to the component-wise updating of regression coefficients in our algorithm.³⁰ A possible solution for this is to sample β simultaneously through the consideration of correlated proposals such as the Metropolis-Hastings algorithm based on the iterative weighted least squares.⁴² This could be an area to be explored in future research.

Acknowledgments

The PFS data was derived based on raw data sets obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the website has contributed to, approved or are in any way responsible for the contents of this publication.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number SC2GM135078.

References

- Law CG and Brookmeyer R. Effects of midpoint imputation on the analysis of doubly censored data. Stat Med 1992; 11: 1569–1578. [PubMed: 1439361]
- 2. Gao F, Zeng D and Lin DY. Semiparametric estimation of the accelerated failure time model with partly interval-censored data. Biometrics 2017; 73: 1161–1168. [PubMed: 28444688]
- 3. Kim JS. Maximum likelihood estimation for the proportional hazards model with partly intervalcensored data. J R Stat Soc Ser B 2003; 65: 489–502.
- Huang J Asymptotic properties of nonparametric estimation based on partly interval-censored data. Stat Sin 1999; 9: 501–519.
- Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. J R Stat Soc Ser B 1976; 38: 290–295.
- Zhao X, Zhao Q, Sun J, et al. Generalized log-rank tests for partly interval-censored failure time data. Biom J 2008; 50: 375–385. [PubMed: 18435504]
- Zhou H and Hanson T. A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially-referenced data. J Am Stat Assoc 2018; 113: 571–581.
- Zhou H and Hanson T. spBayesSurv: Bayesian modeling and analysis of spatially correlated survival data, 2018 URL https://cran.r-project.org/package=spBayesSurv. R package version 1.1.3.
- 9. Komárek A and Lesaffre E. Bayesian accelarated failure time model for correlated interval-censored data with a normal mixture as an error distribution. Stat Sin 2007; 17: 549–569.
- Komárek A bayesSurv: Bayesian survival regression with flexible error and random effects distributions, 2018 URL https://cran.r-project.org/package=bayesSurv. R package version 3.2.
- 11. Pan W Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. J Comput Graph Stat 1999; 8: 109–120.
- Delord M MIICD: multiple imputation for interval censored data, 2016 URL https://CRAN.Rproject.org/package=MIICD. R package version 2.3.
- Reich NG, Lessler J, Cummings D, et al. Estimating incubation period distributions with coarse data. Stat Med 2009; 28: 2769–2784. [PubMed: 19598148]
- 14. Fay MP and Shaw PA. Exact and asymptotic weighted logrank tests for interval censored data: the interval R package. J Stat Softw 2010; 36: 1–38.
- Touraine C, Gerds TA and Joly P. SmoothHazard: an R package for fitting regression models to interval-censored observations of illness-death models. J Stat Softw 2017; 79: 1–22. [PubMed: 30220889]
- Henschel V, Heiss C and Mansmann U. survBayes: fits a proportional hazards model to time to event data by a Bayesian approach, 2012 URL https://CRAN.R-project.org/package=survBayes. R package version 0.2.2.
- Wang W, Chen MH, Wang J, et al. dynsurv: dynamic models for survival data, 2019 URL https:// CRAN.R-project.org/package=dynsurv. R package version 0.3–7.
- Anderson-Bergman C icenReg: regression models for interval censored data, 2019 URL https:// CRAN.R-project.org/package=icenReg. R package version 2.0.13.
- 19. Therneau TM and Lumley T. survival: survival analysis, 2019 URL https://cran.r-project.org/ package=survival. R package version 2.44–1.1.

- Haario H, Saksman E and Tamminen J. An adpative Metropolis algorithm. Bernoulli 2001; 7: 223– 242.
- 21. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970; 57: 97–109.
- 22. Diaconis P and Ylvisaker D. Quantifying prior opinion Technical report no. 207, Stanford University, 10 1983.
- Perron F and Mengersen K. Bayesian nonparametric modeling using mixtures of triangular distributions. Biometrics 2001; 57: 518–528. [PubMed: 11414578]
- 24. Cox DR. Regression models and life-tables (with discussion). J R Stat Soc Ser B 1972; 34: 187–220.
- 25. Cai B, Lin X and Wang L. Bayesian proportional hazards model for current status data with monotone splines. Comput Stat Data Anal 2011; 55: 2644–2651.
- Pan C, Cai B, Wang L, et al. Bayesian semiparametric model for spatially correlated intervalcensored survival data. Comput Stat Data Anal 2014; 74: 198–208.
- Lin X, Cai B, Wang L, et al. A Bayesian proportional hazards model for general interval-censored data. Lifetime Data Anal 2015; 21: 470–490. [PubMed: 25098226]
- 28. Pan C, Cai B and Wang L. Multiple frailty model for clustered interval-censored data with frailty selection. Stat Meth Med Res 2015; 26: 1308–1322.
- Lin X and Wang L. A semiparametric probit model for case 2 interval-censored failure time data. Stat Med 2010; 29: 972–981. [PubMed: 20069532]
- Lin X and Wang L. Bayesian proportional odds models for analyzing current status data: univariate, clustered, and multivariate. Commun Stat Simul Comput 2011; 40: 1171–1181.
- 31. Wang L and Lin X. A Bayesian approach for analyzing case 2 interval-censored failure time data under the semiparametric proportional odds model. Stat Probabil Lett 2011; 81: 876–883.
- 32. Wang L and Dunson DB. Semiparametric Bayes proportional odds models for current status data with under-reporting. Biometrics 2011; 67: 1111–1118. [PubMed: 21175554]
- 33. Ramsay JO. Monotone regression splines in action. Stat Sci 1988; 3: 425–441.
- 34. Geweke J Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Bayesian Stat 1992; 4: 169–193.
- Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. J R Stat Soc Ser B 2002; 64: 583–639.
- 36. Geisser S and Eddy WF. A predictive approach to model selection. J Am Stat Assoc 1979; 74: 153–160.
- Plummer M, Best N, Cowles K, et al. coda: output analysis and diagnostics for MCMC, 2019 URL https://cran.r-project.org/package=coda. R package version 0.19–3.
- 38. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guideline to evaluate the response to treatment of solid tumors. J Natl Cancer Inst 2000; 92: 205–216. [PubMed: 10655437]
- 39. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 53: 457–481.
- 40. Peeters M, Price TJ, Cervantes A, et al. Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. J Clin Oncol 2010; 28: 4706–4713. [PubMed: 20921462]
- 41. Peeters M, Price TJ, Cervantes A, et al. Final results from a randomized phase 3 study of FOLFIRI ± panitumumab for second-line treatment of metastatic colorectal cancer. Ann Oncol 2014; 25: 107–116. [PubMed: 24356622]
- 42. Gamerman D Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing 1997; 7: 57–68.



Figure 1:

Simulation I - Plot of estimated $S_0(t)$ based on 100 simulated data sets using the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph compared to true $S_0(t)$ curve.



Figure 2:

Simulation II - Plot of estimated $S_0(t)$ based on 100 simulated data sets using the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph compared to true $S_0(t)$ curve.

Pan et al.

Page 16



Figure 3:

Metastatic colorectal cancer trial (N = 855) - Estimated survival curves using the proposed method, survregbayes, survregbayes2, bayessurvreg1, coxph, and Kaplan-Meier method. Four curves are plotted for each method based on the four groups formed by treatment arm and mutation status.

Table 1:

Simulation I - Estimation of regression coefficient, effective sample size, absolute value of Geweke's Z-score, deviance information criterion, and nagetive log-likelihood based on the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph.

R function	True	Estimate	SSD	ESE	95CP	ESS	Geweke's Z	DIC	NLLK
Proposed method	1	1.012	0.125	0.130	0.94	557	0.7775	560	297
	1	1.004	0.117	0.121	0.95	834	0.7790		
survregbayes	1	0.993	0.120	0.130	0.97	1099	0.9755	688	344
	1	1.017	0.118	0.124	0.96	1081	0.9323		
survregbayes2	1	0.982	0.118	0.129	0.96	1120	0.8147	689	345
	1	1.006	0.118	0.122	0.95	1109	0.8523		
bayessurvregl	1	-1.379	0.169	0.184	-	1787	1.1591	-	769
	1	-1.443	0.188	0.178	-	1019	1.0351		
coxph	1	0.634	0.116	0.108	0.11	-	-	-	1801
	1	0.702	0.126	0.110	0.23	-	-		

Table 2:

Simulation II - Estimation of regression coefficient, effective sample size, absolute value of Geweke's Z-score, deviance information criterion, and negative log-likelihood based on the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph.

R function	True	Estimate	SSD	ESE	95CP	ESS	Geweke's Z	DIC	NLLK
Proposed method	1	0.982	0.142	0.145	0.96	135	1.1637	422	210
	1	0.989	0.144	0.145	0.94	188	0.9371		
survregbayes	1	0.873	0.122	0.138	0.89	1028	1.1991	431	216
	1	0.884	0.122	0.136	0.87	1022	1.1645		
survregbayes2	1	0.804	0.125	0.138	0.69	997	1.3432	441	223
	1	0.813	0.128	0.135	0.72	952	1.5399		
bayessurvregl	1	-0.499	0.061	0.072	-	1536	0.9370	-	275
	1	-0.504	0.075	0.072	-	784	1.1358		
coxph	1	0.307	0.096	0.100	0	-	-	-	1998
	1	0.324	0.102	0.101	0	-	-		

Table 3:

Metastatic colorectal cancer trial (N = 855) - Estimation of regression coefficient, effective sample size, and negative log-likelihood based on the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph.

R function		Estimate	SE	95% CI	ESS	NLLK
Proposed method	Treatment	-0.229	0.085	(-0.395, -0.062)	2657	1441
	KRAS	0.131	0.086	(-0.037, 0.298)	3165	
survregbayes	Treatment	-0.186	0.085	(-0.355, -0.019)	1294	1557
	KRAS	0.149	0.085	(-0.018, 0.317)	1264	
survregbayes2	Treatment	-0.171	0.088	(-0.341, 0.003)	1158	1564
	KRAS	0.140	0.089	(-0.035, 0.310)	1178	
bayessurvregl	Treatment	0.239	0.094	(0.059, 0.428)	1047	1280
	KRAS	-0.169	0.101	(-0.365, 0.034)	640	
coxph	Treatment	-0.215	0.086	(-0.384, -0.046)	-	3230
	KRAS	0.163	0.086	(-0.006, 0.332)	-	

Table 4:

Metastatic colorectal cancer trial (N = 855) - Estimation of regression coefficient, effective sample size, and negative log-likelihood among patients with wild-type KRAS tumors and patients with mutant KRAS tumors, based on the proposed method, survregbayes, survregbayes2, bayessurvreg1, and coxph.

R function		Estimate	SE	95% CI	ESS	NLLK
Proposed method	wild-type	-0.473	0.108	(-0.685, -0.264)	3220	903
	mutant	-0.260	0.116	(-0.489, -0.032)	3250	663
survregbayes	wild-type	-0.306	0.119	(-0.537, -0.077)	1532	896
	mutant	-0.031	0.132	(-0.285, 0.227)	1591	663
survregbayes2	wild-type	-0.291	0.120	(-0.525, -0.054)	1641	897
	mutant	-0.029	0.131	(-0.289, 0.228)	1675	659
bayessurvregl	wild-type	0.377	0.130	(0.129, 0.636)	1127	744
	mutant	0.059	0.135	(-0.200, 0.325)	675	496
coxph	wild-type	-0.336	0.117	(-0.565, -0.107)	-	1590
	mutant	-0.057	0.128	(-0.306, 0.193)	-	1265