Published in final edited form as:

Biometrics. 2011 September ; 67(3): 1111–1118. doi:10.1111/j.1541-0420.2010.01532.x.

Semiparametric Bayes' Proportional Odds Models for Current Status Data with Underreporting

Lianming Wang^{1,*} and David B. Dunson^{2,**}

¹Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A

²Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A

Summary

Current status data are a type of interval-censored event time data in which all the individuals are either left or right censored. For example, our motivation is drawn from a cross-sectional study, which measured whether or not fibroid onset had occurred by the age of an ultrasound exam for each woman. We propose a semiparametric Bayesian proportional odds model in which the baseline event time distribution is estimated nonparametrically by using adaptive monotone splines in a logistic regression model and the potential risk factors are included in the parametric part of the mean structure. The proposed approach has the advantage of being straightforward to implement using a simple and efficient Gibbs sampler, whereas alternative semiparametric Bayes' event time models encounter problems for current status data. The model is generalized to allow systematic underreporting in a subset of the data, and the methods are applied to an epidemiologic study of uterine fibroids.

Keywords

Cross-sectional; Interval censored; Measurement error; Monotone splines; Proportional odds model; Survival analysis; Uterine fibroids

1. Introduction

Uterine leiomyomata, commonly called fibroids, are the leading cause of hysterectomy for women approaching the age of menopause in the United States (Farquhar and Steiner, 2002). Uterine leiomyomata are found to be associated with adverse pregnancy outcomes, such as difficulty conceiving, preterm birth, and cesarean delivery (Cornado, Marshall, and Schwartz, 2000; Klatsky et al., 2008; Laughlin et al., 2009). Right from the Start (RFTS) is a prospective, cohort study of early pregnancy that includes women from three US states: North Carolina, Tennessee, and Texas. Eligible women were 18 years or older, enrolled by 12 6/7 weeks of gestation based on last menstrual period, did not use assisted reproductive technology, intended to carry the pregnancy to term, spoke English or Spanish, and did not plan to move for the next 18 months. Endovaginal ultrasound was scheduled for all participants as early as possible in the pregnancy, with a goal of the seventh gestational week. Among the 4821 women who enrolled as of December 2007, 4496 women completed

© 2010, The International Biometric Society

*wang99@mailbox.sc.edu.

**dunson@stat.duke.edu

8. Supplementary Materials

The Web Appendix, referenced in Sections 5 and 6, is available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

an ultrasound through which fibroid status was determined. Research interests of this study include estimating the cumulative incidence of fibroids for black and white women and identifying risk factors of fibroids.

RFTS contains three separately funded subsets, RFTS 1, 2, and 3. The investigators reported that there was a possible underreporting problem in RFTS 1 due to less intensive systematic training of the ultrasonographers in RFTS 1. They likely missed some fibroids that would have been detected by the ultrasonographers in RFTS 2 or 3. Table 1 summarizes the total number of women, and the number and the percentage of women who were detected to have fibroids in each substudy for both black and white races studied in Section 6. The percentages in RFTS 2 and RFTS 3 are similar but the percentage in RFTS 1 is much smaller, consistent with the anticipated underreporting problem.

For fibroid data, one never observes the onset time exactly and can only know the fibroid status at the ultrasound examination. Such data are called current status data, also referred to as case 1 interval-censored data. To complete the two research goals mentioned above, we need to conduct regression analysis on current status data. The first research interest corresponds to estimating the cumulative distribution function (CDF) of the fibroid onset time and the second interest corresponds to assessing covariate effects taking the underreporting problem into account.

For analyzing general censored data including current status data, many regression models have been proposed including the proportional hazards (PH) model (Grummer-Strawn, 1993; Huang, 1996; Mongoue-Tchokote and Kim, 2008), the proportional odds (PO) model (Rossini and Tsiatis, 1996; Huang and Rossini, 1997), the additive hazards (Lin, Oakes, and Ying, 1998; Martinussen and Scheike, 2002), and the accelerated failure time model (Betensky, Rabinowitz, and Tsiatis, 2001; Tian and Cai, 2006). These semiparametric models are flexible in allowing an unknown baseline survival function, while maintaining interpretability of the regression parameters characterizing covariate effects.

We focus on the PO models motivated by the close relationship with logistic regression used routinely in analyzing biomedical data. In particular, the regression parameters have a clear interpretation in terms of log odds ratios, and hence are easily interpretable to epidemiologists. We propose a novel Bayesian method under the PO model for analyzing current status data. This leads to computational advantages of using the PO model compared with other semiparametric models for current status data. For example, the accelerated failure time models widely used in the Bayes' semiparametric literature require imputation of the missing event times for tractable posterior computation. We have observed that such imputation leads to major computational problems in current status models, because the event time needs to be imputed for every subject. Bayesian PH and additive hazards are also computationally intensive to implement if one wants to obtain a smooth baseline hazard.

We propose a Bayesian semiparametric PO model for analyzing the fibroid data. Use of a Bayes' approach facilitates adaptive estimation of a smooth baseline survival distribution and extension to accommodate the underreporting problem described above. In addition, the Bayes' approach avoids reliance on large sample assumptions in conducting inferences on the cumulative fibroid incidence curve and covariate effects. We propose to model the unknown baseline log odds function with adaptive monotone splines and estimate the baseline and regression coefficients jointly. An efficient posterior computation procedure is developed based on a data augmentation.

Section 2 provides a review of related methods in the literature and describes our proposed model. Section 3 outlines the posterior computation details. Section 4 generalizes the approach to accommodate underreporting. The proposed methods are evaluated and

compared with a likelihood method through simulation studies in Section 5. Section 6 applies the methods to the fibroid data, and Section 7 provides concluding remarks.

2. Proportional Odds Models for Current Status Data

2.1 Background

Let t_i be the screening or diagnosis time for woman i and y_i be the censoring indicator of onset time of fibroid at t_i . The fibroid onset time is left censored if $y_i = 1$ and right censored if $y_i = 0$. We use $\mathbf{x}_i$ to be the $p \times 1$ covariate vector associated with subject i . The observed data are $\{(y_i, t_i, \mathbf{x}_i), i = 1, \dots, n\}$.

Let $F(\cdot|\mathbf{x})$ and $F_0(\cdot)$ be the CDFs of onset time given covariate $\mathbf{x} = \mathbf{0}$ and $\mathbf{x} = 0$, respectively. We call $F_0(\cdot)$ the baseline CDF of onset time. The PO model specifies that the odds of onset by t given covariate $X = \mathbf{x}$ is proportional to the odds given $X = 0$,

$$\frac{F(t|\mathbf{x})}{1-F(t|\mathbf{x})} = \frac{F_0(t)}{1-F_0(t)} \exp(\mathbf{x}'\boldsymbol{\beta}). \quad (1)$$

Equivalently, we can rewrite the above model as

$$\text{logit } F(t|\mathbf{x}) = \alpha(t) + \mathbf{x}'\boldsymbol{\beta}, \quad (2)$$

where $\alpha(t) = \log [F_0(t)/(1 - F_0(t))]$ is the baseline log odds function that is nondecreasing in t for $F(t)$ to be a valid CDF. The coefficient $\boldsymbol{\beta}_j$ is interpretable as the increase in the log odds of onset by time t attributable to a unit increase in the j th predictor.

Under the independence assumption between the censoring time and onset time, the observed likelihood is

$$L = \prod_{i=1}^n F(t_i|\mathbf{x}_i)^{y_i} \{1 - F(t_i|\mathbf{x}_i)\}^{1-y_i}. \quad (3)$$

The unknown parameters involved in this likelihood are regression parameters $\boldsymbol{\beta}$ and the baseline distribution function F_0 or equivalently the baseline log odds function α . Because F_0 and α have an infinite dimension, estimating these functions is challenging.

There are many frequentist approaches for models (1)–(3). Among others, Grummer-Strawn (1993) used natural cubic splines for the baseline log odds function without incorporating monotonicity; Rossini and Tsiatis (1996) approximated the baseline odds function by a nondecreasing step function; Huang and Rossini (1997) use piecewise linear sieves to approximate the baseline odds function for analyzing general interval-censored data; Rabinowitz, Betensky, and Tsiatis (2000) proposed a computational method to estimate the regression coefficients based on the conditional logistic model for general interval-censored data; Shiboski (1998) proposed a generalized additive model for current status data using isotonic regression for the baseline; and Shen (2000) proposed a random-sieve likelihood approach. Most of the aforementioned papers require intensive and difficult computation.

There are few articles on semiparametric Bayes' approaches for PO models. Banerjee and Dey (2005) used mixtures of Beta distributions to characterize the baseline log odds function in a PO model for spatially correlated survival data. Hanson and Yang (2007) used a mixture of finite Polya trees for the baseline survival function in a PO model. Both of these

approaches require complex and computationally intensive implementations. Banerjee et al. (2007) proposed a class of generalized odds-rate hazards models, which includes PO models as a special case. They used a piecewise constant model for the baseline hazard, so do not allow smooth non-parametric estimation of the hazard or baseline log-odds functions. These Bayes' approaches were applied to right-censored survival data, and the application to current status data is either nontrivial or highly computationally intensive. Although one could use a data augmentation approach, which imputes the left-censored event times within a Markov chain Monte Carlo (MCMC) algorithm, such an approach is subject to slow mixing problems in our experience.

2.2 Proposed Approach

To obtain a smooth nonparametric estimate of the baseline log odds function, we propose to characterize the function using adaptive monotone splines,

$$\alpha(t) = \gamma_0 + \sum_{l=1}^k \gamma_l b_l(t), \quad (4)$$

where $\{b_l\}_{l=1}^k$ are I spline basis functions (Ramsay, 1988), each of which starts at 0 in an initial flat region, is increasing in a mid region, and then plateaus at 1 at higher values. These basis functions are piecewise polynomial and provide great flexibility in modeling local changes. To ensure the monotonicity, we restrict the basis coefficients γ_l to be nonnegative for $l = 1, \dots, k$. Note that γ_0 is unconstrained and its existence is necessary for modeling the baseline log odds function because without γ_0 the linear combination can only model nondecreasing functions starting at zero.

I-splines can be used to approximate any smooth, nondecreasing function as long as sufficient numbers of knots are chosen. However, for finite data sets it is important to avoid choosing too many knots to avoid overfitting. A solution to this problem is to use an adaptive approach, which allows the number and the locations of the knots to be unknown. For example, Shen (1998) proposed a frequentist approach that used monotone splines of a power series form for the baseline function in the PO model for survival data, and Wang (2008) adopted reversible jump Markov chain Monte Carlo (Green, 1995) in monotone curve fitting for complete data. However, these methods are very complicated and computationally expensive.

Instead, Smith and Kohn (1996) choose a large number of potential knots and then use Bayesian variable selection to drop out knots that are not needed. We initially implemented such an approach for knot selection, and obtained good performance in simulations. However, a simpler exponential shrinkage prior with a random mean parameter leads to substantial computational gains without sacrificing estimation performance. Such shrinkage priors tend to favor values close to zero for the unnecessary basis functions, and allow us to use a large number of equally spaced knots to provide sufficient model flexibility while not taking too much computation time.

3. Posterior Computation

We propose a simple data augmentation MCMC algorithm relying on introducing latent variables as follows,

$$\begin{aligned} y_i &= 1_{(z_i > 0)}, \\ z_i &\sim N(\alpha(t_i) + \mathbf{x}_i' \boldsymbol{\beta}, \lambda_i), \\ \lambda_i &= (2\psi_i)^2, \quad \psi_i \sim KS. \end{aligned} \quad (5)$$

In (5), KS denotes the Kolmogorov–Smirnov distribution (Devroye, 1986). If we integrate out all the ψ_i 's, λ_i 's, and z_i 's, we obtain the same form as the observed likelihood in (3). A similar idea is used in Holmes and Held (2006) for standard logistic regression.

Motivated by the widely used double exponential prior in the Bayesian Lasso regression, we assign independent exponential prior $\text{Exp}(\eta)$ for γ_l , $l = 1, \dots, k$, and assign a hyperprior $\mathcal{G}(\mathbf{a}_\eta, b_\eta)$ for η to provide further flexibility. Such prior specification allows borrowing information among γ_l s in updating η and penalizes large values of γ_l s. The same prior specification is also used in Lin and Wang (2010).

We assign a normal distribution $N(m_0, v_0^{-1})$ for γ_0 and a multivariate normal prior $N(\boldsymbol{\beta}_0, \Sigma_0)$ for $\boldsymbol{\beta}$. Both of these priors are natural choices and are conjugate, resulting in easy updating. A data augmentation MCMC algorithm cycles with the following steps:

1. Sample z_i for $i = 1, \dots, n$.

if $y_i = 1$, sample $[z_i | \cdot] \propto N(\alpha(t_i) + \mathbf{x}_i' \boldsymbol{\beta}, \lambda_i) 1_{(z_i > 0)}$.

if $y_i = 0$, $[z_i | \cdot] \propto N(\alpha(t_i) + \mathbf{x}_i' \boldsymbol{\beta}, \lambda_i) 1_{(z_i < 0)}$.

2. Sample γ_0 from $N(E_0, W_0)$, where $W_0 = (v_0 + \sum_{i=1}^n \lambda_i^{-1})^{-1}$, and

$$E_0 = W_0 \left[v_0 m_0 + \sum_{i=1}^n \lambda_i^{-1} \left\{ z_i - \sum_{l=1}^k \gamma_l b_l(t_i) - \mathbf{x}_i' \boldsymbol{\beta} \right\} \right].$$

3. Sample γ_l for $l = 1, \dots, k$. Let $W_l = \sum_{i=1}^n \lambda_i^{-1} b_l^2(t_i)$. If $W_l = 0$, sample γ_l from the prior $\text{Exp}(\eta)$; otherwise, sample γ_l from $N(E_l, W_l^{-1}) 1_{(\gamma_l > 0)}$, where

$$E_l = W_l^{-1} \left[\sum_{i=1}^n \lambda_i^{-1} b_l(t_i) \left\{ z_i - \gamma_0 - \sum_{l' \neq l} \gamma_{l'} b_{l'}(t_i) - \mathbf{x}_i' \boldsymbol{\beta} \right\} - \eta \right].$$

4. Sample $\boldsymbol{\beta}$ from $N(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma})$, where

$$\tilde{\Sigma} = \left(\sum_{i=1}^n \lambda_i^{-1} \mathbf{x}_i \mathbf{x}_i' + \Sigma_0^{-1} \right)^{-1}$$

and

$$\tilde{\boldsymbol{\beta}} = \tilde{\Sigma} \left[\Sigma_0^{-1} \boldsymbol{\beta}_0 + \sum_{i=1}^n \lambda_i^{-1} \{ z_i - \alpha(t_i) \} \mathbf{x}_i \right].$$

5. Sample λ_i using the rejection sampling as in Holmes and Held (2006).
6. Sample η from $\mathcal{G}(a_\eta+k, b_\eta+\sum_{l=1}^k \gamma_l)$.

The above procedure is efficient, with all the full conditionals having a simple closed form except that of λ_i , which can be easily updated using a rejection sampling algorithm. The resulting algorithm is straightforward to implement and has good rates of convergence and mixing in the cases we have considered.

4. Adjusting for Underreporting

As mentioned in Section 1, there is underreporting of fibroids in RFTS 1. We combine RFTS 2 and RFTS 3 together for simplicity of analysis because they were conducted under the same design, appear to have similar data based on simple summary statistics, and RFTS 3 has a small sample size. For woman i ($i = 1, \dots, n$), let $G_i \in \{1, 2\}$ denote the new substudy number, y_i denote a 1/0 indicator that a fibroid was detected in her ultrasound examination, and R_i denote a 1/0 indicator of her true fibroid status at the examination time t_i . Note that the definition of y_i here matches the observed censoring indicator defined in Section 2. Then for women having $G_i = 2$, $R_i = y_i$, whereas for women having $G_i = 1$, R_i is a potential outcome (or counterfactual). For identifiability of the underreporting bias, we assume that there is no systematic difference between the substudies in the true fibroid prevalence. This assumption is in agreement with study investigators' substantive knowledge. Then, we let $\kappa = \Pr(y_i = 0 \mid R_i = 1, G_i = 1)$ denote the probability that a fibroid that would have been observed in the new RFTS 2 was missed in RFTS 1. To complete a Bayesian specification of the model, we let $\kappa \sim \text{Beta}(a_\kappa, b_\kappa)$. To avoid strong subjective opinion, we use $a_\kappa = b_\kappa = 1$, leading to a uniform distribution for κ .

The data augmentation MCMC algorithm of Section 3 can be easily adapted to accommodate the underreporting problem as follows:

1. Set $R_i = y_i$ for $G_i = 2$, and otherwise sample R_i from $\text{Bernoulli}(p_i)$ with

$$p_i = \frac{\kappa \exp(\alpha(t_i) + x_i' \beta)}{1 + \kappa \exp(\alpha(t_i) + x_i' \beta)}.$$

2. Sample z_i as before but replace y_i with R_i .
3. Sample γ, β, λ , and η as in Section 3.
4. Sample κ from $\text{Beta}(a_\kappa + \sum_i R_i (1 - y_i) 1_{(G_i=1)}, b_\kappa + \sum_i R_i y_i 1_{(G_i=1)})$.

5. Simulation Studies

Simulation studies were conducted to evaluate the proposed methods. We generated 100 data sets with sample size $n = 200$ in each data set from the following model

$$\text{logit } F(t|\mathbf{x}) = \alpha(t) + \beta_1 x_1 + \beta_2 x_2,$$

where x_1 is a $N(0, 1)$ random variable and x_2 is a $\text{Bernoulli}(0.5)$ random variable. Two scenarios were considered with different forms of the baseline log odds function: (1) $\alpha(t) = -4 + 2t$ and (2) $\alpha(t) = -3 + 2t^{1/2} + 0.5t^3$. In each scenario, we took the true β_1 to be -1 or 1 , the true β_2 to be -1 or 1 . To specify the monotone spline in modeling $\alpha(t)$, we adopted I

spline basis with degree equal to 4 and took 20 equally spaced knots in (0, 10) for the case of $\alpha(t) = -4 + 2t$ and in (0, 6) for the case of $\alpha(t) = -3 + 2t^{1/2} + 0.5t^3$. We generated censoring time t_i from an exponential distribution with mean 1 and then generate censoring indicator y_i from a Bernoulli distribution with success probability

$$\frac{\exp(\alpha(t_i) + \beta_1 x_{i1} + \beta_2 x_{i2})}{1 + \exp(\alpha(t_i) + \beta_1 x_{i1} + \beta_2 x_{i2})}.$$

To implement the proposed Bayesian methods, we recommend the following standard hyperparameter specification: $m_0 = 0$ and $v_0 = 0.1$, which allows large variation for γ_0 , $a_\eta = b_\eta = 1$, which gives a Gamma prior $\mathcal{G}(1, 1)$ for η , and $\beta_0 = 0$, $\sum_{i=1}^n (X_i' X_i)^{-1}$, a unit information prior for β . For each setup, we implemented the estimation procedure in Section 3 with a total run of 11,000 iterations for each data set and collected every 10th sample after discarding the first 1000 iterations as a burn-in to eliminate the effect of autocorrelation. All the chains passed the convergence test of Geweke (1992) from our observation. Additional MCMC convergence assessments based on the Gelman and Rubin (1992) can be found in the Web Appendix.

We also applied the maximum likelihood method to the same data sets to compare with the proposed Bayesian method. The likelihood method is plausible here because there are only a finite number of parameters in the model due to the use of monotone spline (4). Let $\theta = (\gamma', \beta')'$, where $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_k)'$. Finding the maximum likelihood estimates $\hat{\theta}$ of θ subject to parameter constraints can be easily done by using constrained optimization routines in standard software packages such as the function `fmincon` in Matlab. The variance estimates of $\hat{\theta}$ can be obtained through the diagonal elements of the inverse of the observed information matrix, $\hat{I}(\hat{\theta})^{-1}$. The (s, l) th element of $\hat{I}(\hat{\theta})$ was approximated numerically by

$$\hat{I}(s, l) \approx -h_n^{-2} \{ \log l(\hat{\theta} + h_n \vec{e}_s) - \log l(\hat{\theta} - h_n \vec{e}_s) - \log l(\hat{\theta} + h_n \vec{e}_l) + \log l(\hat{\theta}) \},$$

where $\vec{e}_s$ is a $(1 + k + p)$ -dimensional vector with the s th element equal to 1 and all other elements equal to 0, h_n is a tuning constant with order of $n^{-1/2}$, and $\log l$ is the log likelihood function of (3). This approximation method was also adopted by Zeng, Cai, and Shen (2006), Zeng and Lin (2007), and Lin and Wang (2010), among others. Taking the tuning parameter $h_n = 0.1n^{-1/2}$ or $0.2n^{-1/2}$ did not produce any difference in the variance estimates from our simulation. It should be noted that about 10% of the data sets failed to converge when using `fmincon` in Matlab and the summary of the likelihood method was based on those converged results only.

Table 2 presents the frequentist operating characteristics of the proposed Bayesian and maximum likelihood estimates, with Point denoting the average of the 100 point estimates, ESD the average of the estimated standard deviations, SSD the sample standard deviation of the 100 point estimates, and CP 95 the 95% coverage probability. The point estimates are taken to be posterior means and maximum likelihood estimates, and the coverage probabilities are based on the credible intervals and confidence intervals under the Bayesian and likelihood methods, respectively. From Table 2, the proposed method seems to work very well and outperform the maximum likelihood method with smaller bias in point estimate, smaller variance estimate, and closer coverage probability to the nominal level 0.95.

We also calculate the mean squared errors of the estimates of $\hat{\beta}_1$, $\hat{\beta}_1$, and $\hat{F}_0$ for all the setups under the Bayesian approach and the maximum likelihood method. The results are presented in Tables 2 and 3 of the Web Appendix. Those results clearly suggest that the proposed Bayesian method outperforms the maximum likelihood method in estimating both regression parameters and the baseline CDF.

We also conducted simulations for evaluating the measurement error adjustment method presented in Section 4. We focused on the same setups as before but took the true missing probability κ equal to 20% when generating data. The results are presented in Table 3 and show larger variance estimates and larger bias in point estimates compared to the results in Table 2 as expected. Overall the adjustment method works reasonably well considering the complication of underreporting in the current status data structure. Simulation results with the true missing probability $\kappa = 10\%$ can be found in the Web Appendix.

6. Fibroid Data Analysis

We focus on the black and white women in the analysis, and participants with other races are not considered here. Four covariates are of particular interest to epidemiologists and thus are included in the model: race (black or white), parity status (whether a participant has given birth before), age of menarche (when a participant had her first period), and bmi (body mass index) status. We use $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ to denote the covariate vector for woman i , with $x_{i1} = 1$ indicating that woman i is black, $x_{i2} = 1$ indicating that woman i has given birth before and 0 otherwise, x_{i3} the standardized age of menarche, and $x_{i4} = 1$ indicating that the woman i has a bmi at least 30, which is the cutoff point for obesity. The final data set analyzed contains 824 black women and 2632 white women that have complete covariate information. We analyze the fibroid data under the PO model (2) with monotone spline formulation of $\alpha(t)$ as in (4).

Because the ultrasound was scheduled to target at the seventh week of gestation for each participant and most participants were not aware of their fibroid status before their ultrasound examinations (Laughlin et al., 2009), it is reasonable to believe that the fibroid onset time and the ultrasound examination time are independent. Thus, the observed likelihood in (3) is valid for this data set.

We adopted the same prior specification recommended in Section 5 and assign a uniform distribution to the missing probability κ with $a_\kappa = b_\kappa = 1$. In specifying monotone splines in (4), we took degree equal to 4 and 15 equally spaced knots within the data range (18, 46). We implemented the proposed method with and without adjusting for measurement error. For each case, we ran MCMC with 11,000 iterations and collect every 10th sample to summarize the results after discarding the first 1000 iterations as a burn-in. The MCMC convergence assessment by using Gelman and Rubin diagnostic on additional 10 independent runs can be found in the Web Appendix. The diagnostic results suggest that the chains converge very well.

Table 4 shows the estimates of covariate effects under the proposed method with and without measurement error adjustment. As expected, race has a significant effect on the cumulative incidence of fibroid: the odds of developing fibroids for black women is about $e^{1.48} \approx 4.4$ times as that for white women. The parity status and age of menarche also have a significant negative effect on the cumulative incidence rate. In contrast, obesity is not a risk factor for fibroids in this data set. Figure 1 shows the estimated fibroid cumulative incidence for black and white women as well as their 95% credible intervals controlling all other predictors at 0. Figure 2 shows the difference between the fibroid cumulative incidences for the black and white races and its pointwise 95% credible intervals across the age range. All

the results suggest that black women have a higher risk of developing fibroids at all ages in (18, 46) than white women.

Study investigators are also interested in whether other covariates, especially parity, have a different effect on the fibroid risk between black and white women. To investigate this question, we fit the PO model with all the covariates and their interactions with race included. The results are presented in Table 5. We obtained similar results for each covariate shown in Table 4. The effects of three interactions are not significant because all the 95% credible intervals contain 0. This indicates that there is no significant evidence about the interactions between race and other covariates on fibroid incidence.

To formally conduct model comparison between the two models with and without the interaction terms from a Bayesian perspective, we adopt pseudomarginal likelihood (PML) defined as the summation of log conditional predictive ordinates (CPO) of all the observations (Geisser and Eddy, 1979; Gelfand, Dey, and Chang, 1992; Gelfand and Dey, 1994). Let $CPO_M(i)$ denote the CPO of subject i under the model M representing the posterior predictive probability of the observed data from the i th subject given the rest of the data under model M . Large CPO values indicate strong evidence that the data support the model M at the subject level, and a large value of PML indicates strong overall evidence to support M from all the data. Following Gelfand et al. (1992) and Sinha, Chen, and Ghosh (1999), we have

$$\begin{aligned} CPO(i) &= P(T_i \in (l_i, r_i] | \text{Data}_{(-i)}) \\ &= [E\{P(T_i \in (l_i, r_i] | \beta, \alpha)\}^{-1}]^{-1}, \end{aligned}$$

where the observed interval $(l_i, r_i]$ is $(0, t_i)$ if $y_i = 1$ and is (t_i, ∞) if $y_i = 0$ under the current status data structure, and $\text{Data}_{(-i)}$ denote all data deleting the i th observation. Under the PO models, $E\{P(T_i \in (l_i, r_i] | \beta, \alpha)\}^{-1}$ can be approximated as

$$\frac{1}{C} \sum_{c=1}^C \frac{1 + \exp\{\alpha^{(c)}(t_i) + x_i' \beta^{(c)}\}}{\exp[y_i \{\alpha^{(c)}(t_i) + x_i' \beta^{(c)}\}]},$$

where $\alpha^{(c)}(t) = \gamma_0^{(c)} + \sum_{l=1}^k \gamma_l^{(c)} b_l(t)$, and $\gamma_0^{(c)}, \gamma_l^{(c)}$ s, and $\beta^{(c)}$ are the Gibbs samples at iteration c after burn-in. The obtained PMLs are -1107.8 and -1105.9 for the PO models with and without the interaction terms, respectively. Hence, the model without the interactions is preferred.

To evaluate the influence of the knot placement and degree of the monotone splines, we implemented the proposed method with various degrees and numbers of equally spaced knots. The results are displayed in Table 5 of the Web Appendix and suggest that the proposed method is robust regarding the choice of degree and number of knots when the number is at least moderate.

The use of logistic link function in specifying the relationship between the CDF of failure time and covariates leads to the PO model used in this article, while using the c-log-log link function results in the PH model (Shiboski, 1998). We fitted the fibroid data with the PH model with and without adjusting underreporting problem, and the results can be found in Table 6 of the Web Appendix. Under the PH model, we obtain similar covariate effect

estimates and the same conclusion about the risk factors as under the PO model. This suggests that our analysis results are robust to the choice of the link function.

7. Concluding Remarks

In this article, we propose a fully Bayesian method for analyzing current status data with the PO model. Specifically, we adopt monotone splines to model the nondecreasing baseline log odds function. An efficient Bayesian estimation method is developed relying on a data augmentation. Simulation and application suggest that the proposed approach works well and is robust to the choice of the degree and knot placement of monotone splines. We also solve the underreporting problem under the PO model.

Using monotone splines to model an unknown nondecreasing function reduces the number of parameters from infinite to moderate finite and hence facilitates computation. It also provides a smooth estimate of the baseline survival function as well as its credible intervals under the Bayesian method. Under the same setup, the likelihood method is also easy to implement and runs much faster than the Bayesian method. However, the likelihood method undergoes nonconvergence issue for some data sets, likely due to too many parameters in the model resulting from a large number of knots used. A penalized likelihood approach may be more appropriate in this case (Cai and Betensky, 2003; Ma and Kosorok, 2005).

The proposed method can be generalized to handle more complex models, such as the model with time-dependent or nonlinear predictors, and more complicated data, such as clustered or multivariate current status data and case 2 interval-censored data. More research effort is warranted on such topics from a Bayesian perspective as frequentist methods currently are dominating the literature.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Banerjee S, Dey DK. Semiparametric proportional odds models for spatially correlated survival data. *Lifetime Data Analysis*. 2005; 11:175–191. [PubMed: 15938545]
- Banerjee T, Chen MH, Dey DK, Kim S. Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis*. 2007; 13:241–260. [PubMed: 17401683]
- Betensky RA, Rabinowitz D, Tsiatis AA. Computationally simple accelerated failure time regression for interval censored data. *Biometrika*. 2001; 88:703–711.
- Cai T, Betensky RA. Hazard regression for interval-censored data with penalized spline. *Biometrics*. 2003; 59:570–579. [PubMed: 14601758]
- Cornado GD, Marshall LM, Schwartz SM. Complications in pregnancy, labor, and delivery with uterine leiomyomas: A population-based study. *Obstetrics & Gynecology*. 2000; 95:764–769. [PubMed: 10775744]
- Devroye, L. *Non-Uniformly Random Variate Generation*. New York: Springer-Verlag; 1986.
- Farquhar CM, Steiner CA. Hysterectomy rates in the United States 1990–1997. *Obstetrics & Gynecology*. 2002; 99:229–234. [PubMed: 11814502]
- Geisser S, Eddy W. A predictive approach to model selection. *Journal of the American Statistical Association*. 1979; 74:153–160.
- Gelfand AE, Dey DK. Bayesian model choice: Asymptotics and exact calculations (with discussion). *Journal of the Royal Statistical Society, Series B*. 1994; 56:501–514.
- Gelfand, AE.; Dey, DK.; Chang, H. Model determination using predictive distributions with implementation via sampling-based methods. In: Bergernado, JM.; Berger, JO.; Dawid, AP.;

- Smith, AFM., editors. Bayesian Statistics. Vol. 4. New York: Oxford University Press; 1992. p. 147-168.
- Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7:457–511.
- Geweke, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bergernado, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics. Vol. 4. Oxford: Oxford University Press; 1992. p. 169-194.
- Green PJ. Reversible jump Markow chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 1995; 82:711–732.
- Grummer-Strawn LM. Regression analysis of current-status data: An application to breast-feeding. *Journal of the American Statistical Association*. 1993; 88:758–765. [PubMed: 12155396]
- Hanson T, Yang M. Bayesian semiparametric proportional odds models. *Biometrics*. 2007; 63:88–95. [PubMed: 17447933]
- Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*. 2006; 1:145–168.
- Huang J. Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics*. 1996; 24:540–568.
- Huang J, Rossini JA. Sieve estimation for the proportional odds model with interval-censoring. *Journal of the American Statistical Association*. 1997; 92:960–967.
- Klatsky PC, Tran ND, Caughey AB, Fujimoto VY. Fibroids and reproductive outcomes: A systematic literature review from conception to delivery. *American Journal of Obstetrics & Gynecology*. 2008; 198:357–380. [PubMed: 18395031]
- Laughlin SK, Baird DD, Savitz DA, Herring AH, Hartmann KE. Prevalence of uterine leiomyomas in the first trimester of pregnancy. *Obstetrics & Gynecology*. 2009; 113:630–635. [PubMed: 19300327]
- Lin DY, Oakes D, Ying ZL. Additive hazards regression with current status data. *Biometrika*. 1998; 85:289–298.
- Lin X, Wang L. A semiparametric probit model for case 2 interval-censored failure time data. *Statistics in Medicine*. 2010; 29:972–981. [PubMed: 20069532]
- Ma SG, Kosorok M. Penalized log-likelihood estimation for partly linear transformation models with current status data. *The Annals of Statistics*. 2005; 33:2256–2290.
- Martinussen T, Scheike TH. Efficient estimation in additive hazards regression with current status data. *Biometrika*. 2002; 89:649–658.
- Mongoue-Tchokote S, Kim JS. New statistical software for the proportional hazards model with current status data. *Computational Statistics & Data Analysis*. 2008; 52:4272–4286.
- Rabinowitz D, Betensky RA, Tsiatis AA. Using conditional logistic regression to fit proportional odds models to interval censored data. *Biometrics*. 2000; 56:511–518. [PubMed: 10877311]
- Ramsay JO. Monotone regression splines in action. *Statistical Science*. 1988; 3:425–441.
- Rossini A, Tsiatis AA. A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*. 1996; 91:713–721.
- Shen XT. Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*. 1998; 85:165–177.
- Shen XT. Linear regression with current status data. *Journal of the American Statistical Association*. 2000; 95:842–852.
- Shiboski SC. Generalized additive models for current status data. *Lifetime Data Analysis*. 1998; 4:29–50. [PubMed: 9567054]
- Sinha D, Chen MH, Ghosh SK. Bayesian analysis and model selection for interval-censored survival data. *Biometrics*. 1999; 55:585–590. [PubMed: 11318218]
- Smith M, Kohn R. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*. 1996; 75:317–343.
- Tian L, Cai TX. On the accelerated failure time model for current status and interval censored data. *Biometrika*. 2006; 93:329–342.

- Wang X. Bayesian free-knot monotone cubic spline regression. *Journal of Computational and Graphical Statistics*. 2008; 17:373–387.
- Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B*. 2007; 69:507–564.
- Zeng D, Cai J, Shen Y. Semiparametric additive risks model for interval-censored data. *Statistica Sinica*. 2006; 16:287–302.

Figure 1.

Estimates of the cumulative incidence of fibroids for black and white women with measurement error adjustment. The solid curve is the cumulative incidence estimate and the two dotted curves are the 95% credible intervals for black women. The dashed curve is the cumulative incidence estimate and the two dash-dotted curves are the 95% credible intervals for white women.

Figure 2. Difference of the cumulative incidences of fibroid between black and white races and its pointwise 95% credible intervals.

Table 1
Total number of women, the number of women who had fibroids, and the proportion of women who had fibroids in each substudy

Substudy	Black women			White women		
	Total	Case	Proportion	Total	Case	Proportion
RFTS 1	594	86	14.5%	1192	69	5.8%
RFTS 2	227	67	29.5%	1377	146	10.6%
RFTS 3	3	1	33.3%	63	6	9.5%

Table 2

Simulation results based on 100 data sets in each setup

True	Est	Bayesian method			Likelihood method				
		Point	ESD	SSD	CP95	Point	ESD	SSD	CP95
Scenario 1									
$\beta_1 = 1$	$\hat{\beta}_1$	0.9505	0.2550	0.2437	0.940	1.0733	0.2759	0.3251	0.954
$\beta_2 = 1$	$\hat{\beta}_2$	0.9251	0.4256	0.4314	0.980	1.1369	0.4906	0.5155	0.943
$\beta_1 = 1$	$\hat{\beta}_1$	0.9272	0.2615	0.2817	0.970	1.0996	0.3332	0.3715	0.942
$\beta_2 = -1$	$\hat{\beta}_2$	-0.9409	0.5147	0.5133	0.930	-1.1556	0.6115	0.6357	0.919
$\beta_1 = -1$	$\hat{\beta}_1$	-0.9495	0.2570	0.2421	0.960	1.0604	0.2710	0.3561	0.954
$\beta_2 = 1$	$\hat{\beta}_2$	0.9021	0.4012	0.4277	0.950	1.0892	0.4832	0.4885	0.930
$\beta_1 = -1$	$\hat{\beta}_1$	-0.9169	0.2755	0.2795	0.940	-1.1052	0.3374	0.3743	0.933
$\beta_2 = -1$	$\hat{\beta}_2$	-0.9326	0.4767	0.5141	0.970	-1.0998	0.6163	0.6288	0.956
Scenario 2									
$\beta_1 = 1$	$\hat{\beta}_1$	1.0302	0.2195	0.2259	0.950	1.1096	0.2467	0.2609	0.907
$\beta_2 = 1$	$\hat{\beta}_2$	0.9923	0.3713	0.3854	0.970	1.1290	0.4225	0.4511	0.930
$\beta_1 = 1$	$\hat{\beta}_1$	1.0022	0.2710	0.2442	0.940	1.1252	0.2744	0.3570	0.914
$\beta_2 = -1$	$\hat{\beta}_2$	-0.9601	0.4143	0.4318	0.940	-1.0519	0.4795	0.5082	0.936
$\beta_1 = -1$	$\hat{\beta}_1$	-0.9699	0.2236	0.2219	0.930	-1.0428	0.2402	0.2623	0.945
$\beta_2 = 1$	$\hat{\beta}_2$	0.9573	0.3769	0.3828	0.970	1.0749	0.4145	0.4261	0.945
$\beta_1 = -1$	$\hat{\beta}_1$	-0.9711	0.2064	0.2479	0.970	-1.0802	0.2747	0.2499	0.978
$\beta_2 = -1$	$\hat{\beta}_2$	-0.9733	0.3676	0.4360	0.980	-1.0972	0.4832	0.4005	0.989

Table 3
Simulation results of the proposed Bayesian method with measurement error adjustment

True	Est	$\alpha(t) = -4 + 2t$				$\alpha(t) = -3 + 2t^{1/2} + 0.5t^3$			
		Point	ESD	SSD	CP95	Point	ESD	SSD	CP95
$\beta_1 = 1$	$\hat{\beta}_1$	1.0034	0.2949	0.3274	0.960	0.9853	0.2725	0.2915	0.940
$\beta_2 = 1$	$\hat{\beta}_2$	0.9452	0.5088	0.5475	0.960	0.9333	0.4161	0.4721	0.960
$\beta_1 = 1$	$\hat{\beta}_1$	1.0368	0.3445	0.3772	0.960	0.9267	0.2837	0.2934	0.940
$\beta_2 = -1$	$\hat{\beta}_2$	-1.1834	0.5611	0.6439	0.990	-0.9046	0.4609	0.5061	0.950
$\beta_1 = -1$	$\hat{\beta}_1$	-0.9483	0.2694	0.2863	0.970	0.9415	0.4177	0.4655	0.970
$\beta_2 = 1$	$\hat{\beta}_2$	0.9343	0.4930	0.5487	0.960	0.8773	0.4537	0.4663	0.930
$\beta_1 = -1$	$\hat{\beta}_1$	-1.0390	0.3614	0.3896	0.950	-0.8939	0.2606	0.2946	0.930
$\beta_2 = -1$	$\hat{\beta}_2$	-1.2602	0.5917	0.6718	0.950	-0.9388	0.4001	0.5075	0.990

Table 4
Estimated covariate effects on the fibroid incidence: posterior means and 95% credible intervals

Risk factors	Before adjustment		After adjustment	
	Mean	95% CI	Mean	95% CI
Race	1.3815	(1.1328, 1.6371)	1.4714	(1.1781, 1.7820)
Parity	-0.4201	(-0.6486, -0.1818)	-0.5400	(-0.7816, -0.2854)
Menarche	-0.1906	(-0.3018, -0.0765)	-0.1878	(-0.3080, -0.0685)
bmi	0.0562	(-0.2267, 0.3196)	0.0368	(-0.2581, 0.3378)

Table 5

Estimated effects of covariates and their interactions with race on the fibroid incidence with measurement error adjustment

Risk factors	Mean	95% CI
Race	1.4895	(0.9907, 2.2312)
Parity	-0.4897	(-0.7926, -0.0652)
menarche	-0.1950	(-0.3518, -0.0449)
bmi	-0.1696	(-0.6085, 0.2931)
Race * parity	-0.1501	(-0.8670, 0.4105)
Race * menarche	-0.0079	(-0.2715, 0.2431)
Race * bmi	0.4175	(-0.2463, 1.0367)