Published in final edited form as:

Biometrics. 2016 March ; 72(1): 222–231. doi:10.1111/biom.12389.

A Flexible, Computationally Efficient Method for Fitting the Proportional Hazards Model to Interval-Censored Data

Lianming Wang[1], **Christopher S. McMahan**[2], **Michael G. Hudgens**[3], and **Zaina P. Qureshi**[4]

Christopher S. McMahan: mcmaha2@clemson.edu

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

[2]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A

[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A

[4]Department of Health Services Policy and Management, University of South Carolina, Columbia, SC 29208, U.S.A

Summary

The proportional hazards model (PH) is currently the most popular regression model for analyzing time-to-event data. Despite its popularity, the analysis of interval-censored data under the PH model can be challenging using many available techniques. This paper presents a new method for analyzing interval-censored data under the PH model. The proposed approach uses a monotone spline representation to approximate the unknown nondecreasing cumulative baseline hazard function. Formulating the PH model in this fashion results in a finite number of parameters to estimate while maintaining substantial modeling flexibility. A novel expectation-maximization (EM) algorithm is developed for finding the maximum likelihood estimates of the parameters. The derivation of the EM algorithm relies on a two-stage data augmentation involving latent Poisson random variables. The resulting algorithm is easy to implement, robust to initialization, enjoys quick convergence, and provides closed-form variance estimates. The performance of the proposed regression methodology is evaluated through a simulation study, and is further illustrated using data from a large population-based randomized trial designed and sponsored by the United States National Cancer Institute.

Keywords

EM algorithm; interval-censored data; latent Poisson random variables; monotone splines; proportional hazards model

Correspondence to: Christopher S. McMahan, mcmaha2@clemson.edu.

Supplementary Material: The Web Appendices, Tables, and Figures referenced in Sections 2-4 are available with this paper at the *Biometrics* website on Wiley Online Library.

1. Introduction

Originally proposed by Cox (1972), the proportional hazards (PH) model has been widely used for the purposes of analyzing time-to-event data, with its gain in popularity being attributed to its interpretability and ability to model right-censored data. Unfortunately, the development of techniques that allow for the analysis of interval-censored data under semiparametric variants of this model can prove to be quite challenging. These difficulties are encountered because of the underlying structure of interval-censored data; i.e., the event times of interest are never observed. In particular, data of this form typically consist of left-, right-, and interval-censored observations corresponding to the situation in which the event times occur before the first, after the last, or between two observation times, respectively. Interval-censored data is ubiquitous among social, behavioral, epidemiological, and medical studies (Sun, 2006), and therefore modeling techniques that allow for the valid analysis of interval-censored data need to be developed, along with the necessary statistical software to carry out these analyses.

The regression analysis of interval-censored data under the PH model is a well studied problem. This problem was first addressed by Finkelstein (1986), who proposed a method of jointly estimating the regression parameters and the baseline hazard function using a Newton-Raphson based algorithm. Focusing solely on the estimation of the regression parameters, Satten (1996) proposed a marginal likelihood approach and Goggins et al. (1998) developed a Monte Carlo expectation maximization algorithm. Even though these methods avoid estimating the baseline hazard function they remain computationally expensive because they require the imputation of all possible rankings of the failure times that are consistent to the observed data. For interval-censored data without covariates, Turnbull (1976) developed an algorithm based on the idea of self consistency, Groeneboom and Wellner (1992) presented an iterative convex minorant algorithm, and Zhang and Jamshidian (2004) proposed a generalization of the Rosen algorithm (Rosen, 1960) for efficiently computing the nonparametric maximum likelihood estimate of the distribution function of the event/failure time. Pan (1999) reformulated the iterative convex minorant algorithm proposed by Groeneboom and Wellner (1992) as a generalized gradient projection method which allowed for the incorporation of regression parameters. Pan (2000) developed a semiparametric alternative, based on multiple imputation, to the existing nonparametric methods. Goeteghebeur and Ryan (2000) developed an expectation-maximization (EM) algorithm with an M-step that updates estimates of the regression parameters by maximizing a Cox partial likelihood and estimates the baseline hazard function using the Breslow estimator. Betensky et al. (2002) presented local likelihood techniques for fitting the PH model which results in a smooth and interpretable estimate of the baseline hazard as well as an assessment of global covariate effects. Using penalized likelihood methods Cai and Betensky (2003) modeled the log-hazard function with a piecewise-linear spline. Zhang, Hua, and Huang (2010) extended the earlier work of Zhang and Jamshidian (2004) by allowing for covariate effects and by using monotone B-splines to model the cumulative baseline hazard function. Shao et al. (2014) proposed a semiparametric varying-coefficient model for interval-censored data with a cured proportion. For a comprehensive review of the

recent work relating to the analysis of interval-censored data, see Sun (2006), Zhang and Sun (2010), and Li and Ma (2013).

The vast majority of the aforementioned work can be either too computationally intensive or complex for practitioners to implement. Consequently, many study investigators tend to ignore interval-censoring and instead opt to use the midpoint or the right end point of the observed interval as the exact failure time for those left- and interval-censored observations and then adopt the well-established partial likelihood method using `coxph` in R or `PHREG` in SAS (Gómez et al. 2009; Allison, 2010). Though common, this approach may result in biased estimation and inference as has been demonstrated by Rucker and Messerer (1988), Odell, Anderson, and D'Agostino (1992), among many others.

Most existing statistical packages that conduct regression analysis of interval-censored data primarily focus on parametric models, such as `LIFEREG` in SAS and `survreg` in R. To date there exist only a few publicly available packages that perform semiparametric analysis of interval-censored data. The R package `intcox` (Henschel and Mansmann, 2013) adopts the generalized gradient projection method of Pan (1999), but does not provide variance estimates and often obtains biased parameter estimates (Pan 1999; Gómez et al. 2009).

Given the omnipresent nature of interval-censored data, there exists a pressing need to develop flexible, accurate, computationally efficient, and easy-to-implement statistical methods for regression analysis of data of this form. To this end, a new method for analyzing interval-censored data under the PH model is presented herein. The proposed approach meets all of the aforementioned criteria. The methodological details of the proposed technique are provided in Section 2. These details include the use of monotone splines for approximating the cumulative baseline hazards function in the PH model, a two-stage data augmentation process that leads to the development of an EM algorithm that can be used to find the maximum likelihood estimates of all unknown parameters, and closed-form expressions of the asymptotic variance estimates. The performance of the proposed approach is illustrated in Section 3 through an extensive simulation study. In Section 4 the proposed method is used to analyze data from a large population-based randomized trial designed and sponsored by the United States National Cancer Institute. Section 5 provides a summary discussion. As a companion to this work, an R package that implements the proposed methodology has been developed and is freely available from the Comprehensive R Archive Network (CRAN).

2. The proposed method

2.1 Data, model, and observed likelihood

Let $F(\cdot|\boldsymbol{x})$ denote the cumulative distribution function (CDF) of the event/failure time of interest given the covariate vector $\boldsymbol{x}$. Under the PH model the failure time distribution for individuals with covariates $\boldsymbol{x}_i$ is given by $F(t|\boldsymbol{x}_i)=1-\exp\{-\Lambda_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}$, where $\boldsymbol{x}_i=(x_{i1}, \ldots, x_{ip})'$ is a $p\times 1$ vector of time-independent covariates, $\boldsymbol{\beta}=(\beta_1, \ldots, \beta_p)'$ is the corresponding vector of regression parameters, and $\Lambda_0(\cdot)$ is the cumulative baseline hazard function. It is assumed throughout that, conditional on the covariates, the failure time is independent of the observational process. This assumption is common in the survival

literature; see e.g., Liu and Shen (2009) and Zhang and Sun (2010) among others. Under this assumption, the likelihood given the observed data $\mathscr{D}=\{(L_i, R_i], \boldsymbol{x}_i\}_{i=1}^n$ is

$$L_{\text{obs}}=\prod_{i=1}^{n}\{F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i)\},$$

where L_i and R_i denote the left and right bounds of the observed interval for the ith subject, respectively, with $L_i < R_i$. Note, $L_i = 0$ ($R_i = \infty$) indicates that the ith subject's failure time is left (right) censored. Distinguishing between the three types of censoring, one can rewrite the observed data likelihood in the following form

$$L_{\text{obs}}=\prod_{i=1}^{n}F(R_i|\boldsymbol{x}_i)^{\delta_{i1}}\{F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i)\}^{\delta_{i2}}\{1 - F(L_i|\boldsymbol{x}_i)\}^{\delta_{i3}}, \quad (1)$$

where δ_{i1}, δ_{i2}, and δ_{i3} are censoring indicators for the ith subject denoting left-, interval-, and right-censoring, respectively, subject to the constraint $\delta_{i1} + \delta_{i2} + \delta_{i3} = 1$.

2.2 Modeling $\Lambda_0(\cdot)$ with monotone splines

The unknown parameters in the observed data likelihood include the regression coefficients and the cumulative baseline hazard function. It is well known that for right-censored data, partial likelihood methods allow one to consistently estimate $\boldsymbol{\beta}$, without having to estimate $\Lambda_0(\cdot)$, under the PH model. However, partial likelihood techniques are not well suited for interval-censored data. Moreover, the use of counting processes and martingale theory, which work elegantly for right-censored data, do not appear to be directly applicable in the analysis of interval-censored data due to its complex structure (Sun, 2006).

Estimating $\Lambda_0(\cdot)$ is challenging from both a theoretical and computational perspective because of its infinite dimension. To reduce the number of unknown parameters which need to be estimated while also maintaining adequate modeling flexibility, in this paper $\Lambda_0(\cdot)$ is modeled using I-splines (Ramsay, 1988), following the earlier work of Cai, Lin, and Wang (2011) and McMahan, Wang, and Tebbs (2013). This approach leads to the following representation

$$\Lambda_0(\cdot)=\sum_{l=1}^{k}\gamma_l b_l(\cdot), \quad (2)$$

where the $b_l(\cdot)$'s are integrated spline basis functions, each of which is nondecreasing from 0 to 1, and the γ_l's are nonnegative coefficients which ensure that $\Lambda_0(\cdot)$ is nondecreasing.

To construct the integrated spline basis functions, one needs to specify the degree of the basis splines and choose an increasing sequence of knots within a time range (Ramsay, 1988). The degree controls the overall smoothness of the basis functions; e.g., specifying degree to be 1, 2, or 3 corresponds to the use of linear, quadratic, or cubic basis functions,

respectively. The placement of knots determines the overall modeling flexibility, with more knots in a region equating to greater modeling flexibility in that region. Once the degree and placement of knots are specified, the k spline basis functions are fully determined, where k is equal to the number of interior knots plus the degree (Ramsay, 1988). The calculation of these basis functions is a simple task and an R function is available in the companion R package (see Section 5 below).

In general, the specification of the degree and knot placement has the potential to influence parameter estimation, more so for the former rather than the latter. Larger knot sets generally results in attaining more modeling flexibility at the cost of additional computational burdens and potential over-fitting problems; for further discussion see Cai et al. (2011) and Lin and Wang (2010). Ramsay (1988) recommended using a small number of strategically placed interior knots, e.g., placing knots at the median or quartiles. Using penalized Bayesian methods, Lin and Wang (2010), Wang and Dunson (2011), and Wang and Lin (2011) recommended using approximately 10∼30 equally spaced knots for their application of monotone splines under various survival models for analyzing interval-censored data. When the observation times are sparse in certain regions of the observed time range, the former strategy may be more appropriate when compared to the latter, but the findings presented herein suggest that both knot placement schemes perform well in application; e.g., see Sections 3 and 4. Consequently, following the recommendations of the aforementioned authors, one could use either equally spaced knots within the observed time range or knots placed at the quantiles of the finite end points of the observed intervals. Further, adequate smoothness can usually be attained by specifying the degree of the basis splines to be either 2 or 3. For a particular data set, the proposed method should be applied with several different knot placement schemes, to include varying the number of knots, thus resulting in several candidate models. The final model can then be chosen according to a model selection criteria, e.g., Akaike's information criterion (AIC). Similar strategies for determining knot placement are commonly used in the literature; e.g., see Rosenberg (1995) and McMahan et al. (2013).

2.3 Data augmentation for the EM algorithm

Section 2.4 presents an EM algorithm that can be used to find the maximum likelihood estimate of θ, where $\theta = (\beta', \gamma')$ and $\gamma = (\gamma_1, \ldots, \gamma_k)'$. The derivation of the algorithm is based on a two-stage data augmentation involving latent Poisson random variables that exploits the relationship between the PH model and a nonhomogeneous Poisson process.

In what follows, motivation and justification for the proposed data augmentation is provided. Consider a nonhomogeneous Poisson process $N(t)$ having a cumulative intensity function $\Lambda_0(t) \exp(x'\beta)$. Let T denote the time of the first jump of the counting process; i.e., $T = \inf\{t : N(t) > 0\}$. To show that T indeed follows the PH model with a cumulative distribution function $F(t|x) = 1 - \exp\{-\Lambda_0(t) \exp(x'\beta)\}$, note for any t that $\text{pr}(T > t) = \text{pr}\{N(t) = 0\} = \exp\{-\Lambda_0(t) \exp(x'\beta)\} = 1 - F(t|x)$ because $N(t)$ is a Poisson random variable with mean parameter $\Lambda_0(t) \exp(x'\beta)$. Using this relationship, an augmented data likelihood is constructed below, using a latent nonhomogeneous Poisson process.

Let $N_i(t)$ denote the latent Poisson process for subject i, which has cumulative intensity function $\Lambda_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, for $i = 1, \ldots, n$. Define $Z_i = N_i(t_{i1})$, where $t_{i1} = R_i 1_{(\delta i1=1)} + L_i 1_{(\delta i1=0)}$. Similarly, when $\delta_{i1} = 0$ define $W_i = N_i(t_{i2}) - N_i(t_{i1})$, where $t_{i2} = R_i 1_{(\delta_{i2}=1)} + L_i 1_{(\delta i3=1)}$. Thus, Z_i and W_i are Poisson random variables with mean parameters $\Lambda_0(t_{i1})\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ and $\{\Lambda_0(t_{i2}) - \Lambda_0(t_{i1})\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, respectively. Further, note that Z_i and W_i are independent when $\delta_{i1} = 0$. Under this construction, if T_i is left-censored then $\mathrm{pr}(T_i \le t_{i1}) = \mathrm{pr}(N_i(t_{i1}) > 0) = \mathrm{pr}(Z_i > 0) = 1 - \exp\{-\Lambda_0(R_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\} = F(R_i|\boldsymbol{x}_i)$. If T_i is interval-censored, $\mathrm{pr}(t_{i1} < T_i \le t_{i2}) = \mathrm{pr}\{N_i(t_{i1}) = 0, N_i(t_{i2}) > 0\} = \mathrm{pr}(Z_i = 0, W_i > 0) = F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i)$. Similarly, it is easy to show in the case of right-censoring that

$$\mathrm{pr}(T_i > t_{i2}) = \mathrm{pr}\{N_i(t_{i2}) = 0\} = \mathrm{pr}(Z_i = 0, W_i = 0) = 1 - F(L_i|\boldsymbol{x}_i).$$

Based on the latent variables Z_i and W_i, the augmented likelihood can be expressed as

$$L_{\mathrm{aug}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathscr{P}_{Z_i}(Z_i)\mathscr{P}_{W_i}(W_i)^{\delta_{i2}+\delta_{i3}}\{\delta_{i1}1_{(Z_i>0)} + \delta_{i2}1_{(Z_i=0,W_i>0)} + \delta_{i3}1_{(Z_i=0,W_i=o)}\}, \quad (3)$$

where $\mathscr{P}_A(\cdot)$ denotes the probability mass function associated with the random variable A. It is easy to see that one can obtain (1) by integrating the Z_i's and W_i's out of (3).

To exploit the monotone spline representation of $\Lambda_0(\cdot)$ in (2), a second stage of data augmentation is considered. In particular, for each i, both Z_i and W_i are decomposed as sum of k independent Poisson random variables, $Z_i = \sum_{l=1}^{k} Z_{il}$ and $W_i = \sum_{l=1}^{k} W_{il}$, where Z_{il} and W_{il}, for $l = 1, \ldots, k$, are Poisson random variables having mean parameters $\gamma_l b_l(t_{i1})\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ and $\gamma_l\{b_l(t_{i2}) - b_l(t_{i1})\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, respectively. The augmented likelihood associated with the second level of data augmentation is given by

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{l=1}^{k} \mathscr{P}_{Z_{il}}(Z_{il})\mathscr{P}_{W_{il}}(W_{il})^{\delta_{i2}+\delta_{i3}}, \quad (4)$$

where $Z_i > 0$ if $\delta_{i1} = 1$, $Z_i = 0$ and $W_i > 0$ if $\delta_{i2} = 1$, $Z_i = 0$ and $W_i = 0$ if $\delta_{i3} = 1$, $\sum_{l=1}^{k} Z_{il} = Z_i$, and $\sum_{l=1}^{k} W_{il} = W_i$. Again, it is relatively easy to see that by integrating the Z_{il}'s and W_{il}'s out of (4) one can obtain (3). Consequently, the augmented data likelihood (4) can be viewed as the complete data likelihood with all the Z_i's, W_i's, Z_{il}'s, and W_{il}'s being regarded as missing data.

2.4 The EM algorithm

The derivation of the EM algorithm begins by considering the expectation of the logarithm of the complete data likelihood (4) with respect to the latent variables (Z_i's, Z_{il}'s, W_i's, and

W_{il}'s) conditional on the observed data $\mathcal{D}$ and the current parameter estimate $\boldsymbol{\theta}^{(d)} = (\boldsymbol{\beta}^{(d)'}, \boldsymbol{\gamma}^{(d)'})'$. This yields $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) = E[\log\{L_C(\boldsymbol{\theta})\} | \mathcal{D}, \boldsymbol{\theta}^{(d)}]$, which can be expressed as

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) \\ = \sum_{i=1}^{n}\sum_{l=1}^{k} \Big[\{E(Z_{il}|\mathscr{D}, \boldsymbol{\theta}^{(d)}) + (\delta_{i2}+\delta_{i3})E(W_{il}|\mathscr{D}, \boldsymbol{\theta}^{(d)})\}\{\log(\gamma_l) + \boldsymbol{x}_i'\boldsymbol{\beta}\} - \gamma_l \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\{(\delta_{i2}+\delta_{i1})b_l(R_i) + \delta_{i3}b_l(L_i)\} \Big] \\ + L(\boldsymbol{\theta}^{(d)}), \end{aligned} \quad (5)$$

where $L(\boldsymbol{\theta}^{(d)})$ is a function of $\boldsymbol{\theta}^{(d)}$ but is free of $\boldsymbol{\theta}$. The derivation of (5) is provided in Web Appendix A. Noting that the conditional distribution of Z_{il} (W_{il}) given Z_i (W_i) is binomial, and by applying the law of iterated expectations, one can obtain the following conditional expectations

$$\begin{aligned} E(Z_{il}|\mathscr{D}, \boldsymbol{\theta}^{(d)}) &= \{\Lambda_0^{(d)}(R_i)\}^{-1} \gamma_l^{(d)} b_l(R_i) E(Z_i|\mathscr{D}, \boldsymbol{\theta}^{(d)}), \\ E(W_{il}|\mathscr{D}, \boldsymbol{\theta}^{(d)}) &= \{\Lambda_0^{(d)}(R_i) - \Lambda_0^{(d)}(L_i)\}^{-1} \gamma_l^{(d)} \{b_l(R_i) - b_l(L_i)\} E(W_i|\mathscr{D}, \boldsymbol{\theta}^{(d)}), \end{aligned}$$

where $\Lambda_0^{(d)}(\cdot) = \sum_{l=1}^{k} \gamma_l^{(d)} b_l(\cdot)$. Similarly, it can be shown based on the augmented likelihood (3) that Z_i (W_i) conditionally follows a truncated Poisson distribution given the observed data. Therefore, the expected values of Z_i and W_i, given $\mathcal{D}$ and $\boldsymbol{\theta}^{(d)}$, can be expressed as

$$\begin{aligned} E(Z_i|\mathscr{D}, \boldsymbol{\theta}^{(d)}) &= \frac{\Lambda_0^{(d)}(R_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}^{(d)})\delta_{i1}}{1-\exp\{-\Lambda_0^{(d)}(R_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}^{(d)})\}}, \\ E(W_i|\mathscr{D}, \boldsymbol{\theta}^{(d)}) &= \frac{\{\Lambda_0^{(d)}(R_i) - \Lambda_0^{(d)}(L_i)\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}^{(d)})\delta_{i2}}{1-\exp[-\{\Lambda_0^{(d)}(R_i) - \Lambda_0^{(d)}(L_i)\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}^{(d)})]}. \end{aligned}$$

Note $\delta_{i3}E(W_{il}|\mathcal{D}, \boldsymbol{\theta}^{(d)}) = 0$ for all i and l, and these terms are therefore ignored henceforth.

The next step in the EM algorithm finds $\boldsymbol{\theta}^{(d+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$. To this end, consider the partial derivatives of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ with respect to $\boldsymbol{\theta}$, which are given by

$$\begin{aligned} \frac{\partial Q}{\partial \gamma_l} &= \sum_{i=1}^{n} \Big[\gamma_l^{-1}\{E(Z_{il}|\mathscr{D}, \boldsymbol{\theta}^{(d)}) + \delta_{i2}E(W_{il}|\mathscr{D}, \boldsymbol{\theta}^{(d)})\} - \{(\delta_{i1}+\delta_{i2})b_l(R_i) + \delta_{i3}b_l(L_i)\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \Big] \\ \frac{\partial Q}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \Big[\{E(Z_i|\mathscr{D}, \boldsymbol{\theta}^{(d)}) + \delta_{i2}E(W_i|\mathscr{D}, \boldsymbol{\theta}^{(d)})\} - \{(\delta_{i2}+\delta_{i1})\Lambda_0(R_i) + \delta_{i3}\Lambda_0(L_i)\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \Big] \boldsymbol{x}_i. \end{aligned}$$

Clearly, $\boldsymbol{\theta}^{(d+1)}$ is a solution to the system of equations given by $\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})/\partial \boldsymbol{\beta} = 0$ and $\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})/\partial \gamma_l = 0$, for $l = 1, \ldots, k$. By directly solving $\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})/\partial \gamma_l = 0$ for γ_l, a closed-form expression for $\gamma_l^{(d+1)}$ in terms of $\boldsymbol{\beta}^{(d+1)}$ and the observed data for each l can be obtained. Thus, by replacing γ_l in $\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})/\partial \boldsymbol{\beta} = 0$ by the expression for $\gamma_l^{(d+1)}$, for $l = 1, \ldots, k$, and solving for $\boldsymbol{\beta}$ one can obtain $\boldsymbol{\beta}^{(d+1)}$, which then allows for the direct calculation of $\boldsymbol{\gamma}^{(d+1)}$.

In what follows, the EM algorithm is succinctly summarized. First set $d = 0$ and initialize $\boldsymbol{\theta}^{(d)} = (\boldsymbol{\beta}^{(d)\prime}, \boldsymbol{\gamma}^{(d)\prime})'$. Then repeat the following two steps until convergence:

1. Obtain $\boldsymbol{\beta}^{(d+1)}$ by solving the following system of p equations

$$\sum_{i=1}^{n}\{E(Z_i|\mathscr{D},\boldsymbol{\theta}^{(d)})+\delta_{i2}E(W_i|\mathscr{D},\boldsymbol{\theta}^{(d)})\}\boldsymbol{x}_i=\sum_{i=1}^{n}\sum_{l=1}^{k}\{(\delta_{i1}+\delta_{i2})b_l(R_i)+\delta_{i3}b_l(L_i)\}\gamma_l^{*(d)}(\boldsymbol{\beta})\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i,$$

Where

$$\gamma_l^{*(d)}(\boldsymbol{\beta})=\frac{\sum_{i=1}^{n}\{E(Z_{il}|\mathscr{D},\boldsymbol{\theta}^{(d)})+\delta_{i2}E(W_{il}|\mathscr{D},\boldsymbol{\theta}^{(d)})\}}{\sum_{i=1}^{n}\{(\delta_{i1}+\delta_{i2})b_l(R_i)+\delta_{i3}b_l(L_i)\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}.$$

2. Let $\gamma_l^{(d+1)}=\gamma_l^{*(d)}(\boldsymbol{\beta}^{(d+1)})$ and increase d by 1.

Solving the system of equations in the first step of the algorithm can be accomplished using standard root finding routines, available in practically all existing statistical software packages. The second step of the algorithm is a simple updating of $\boldsymbol{\gamma}^{(d)}$ in closed form. Thus, the implementation of the EM algorithm is straightforward and computationally inexpensive. Moreover, it can be shown that $\boldsymbol{\theta}^{(d+1)}$ is the unique global maximizer of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$; a proof of this assertion is provided in Web Appendix B. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})'$ denote the value of $\boldsymbol{\theta}^{(d)}$ at convergence of the EM algorithm. It can be shown that $\hat{\boldsymbol{\theta}}$ solves the score equations based on the observed likelihood (1).

2.5 Asymptotic properties and variance estimation

In this section the asymptotic properties of the proposed estimator are discussed. These properties could be studied under two different assumptions: (S1) the number and position of the knots are known a priori and do not depend on the sample size n; or (S2) that the cardinality of the knot set grows with the sample size (as in Zhang et al. 2010). Proceeding under (S1) implicitly implies that $\Lambda_0(\cdot)$ can be expressed as a linear combination of integrated spline basis functions, whereas (S2) allows for the consistent estimation of $\Lambda_0(\cdot)$ under less stringent assumptions. Under (S1) the general theory of maximum likelihood estimation provides, under the standard regularity conditions, that, as $n \to \infty$,

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N\{0, \mathscr{I}^{-1}(\boldsymbol{\theta})\},$$

where $\mathcal{I}(\boldsymbol{\theta})$ denotes the usual Fisher information matrix. This result holds under the assumption that $\Lambda_0(\cdot)$ can be expressed as (2). If this assumption is in fact invalid, an asymptotic bias may be introduced, although through numerical studies it appears that this bias can be attenuated, and often rendered negligible, when an adequate number of knots is used, e.g., see Section 3.

To derive an estimator of $\mathcal{I}^{-1}(\boldsymbol{\theta})$, Louis's method (Louis, 1982) is adopted. The estimated variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is subsequently given by $I^{-1}(\hat{\boldsymbol{\theta}})$, where

$$I(\boldsymbol{\theta}) = -\frac{\partial^2 Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} - \text{var}\left\{ \frac{\partial \log L_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \mathcal{D}, \hat{\boldsymbol{\theta}} \right\}. \quad (6)$$

The details pertaining to the calculation of the two terms on the right hand side of (6) are provided, in closed-form, in Web Appendix C. These expressions make the variance estimates easy to compute, which is another appealing characteristic of the proposed approach.

3. Simulation study

A series of simulation studies were conducted to assess the characteristics and performance of the proposed methodology across a variety of settings. In particular, three studies were performed which considered (I) low, (II) high, and (III) medium right-censoring rates.

3.1 Simulation study I

This study considers the following true distribution of the failure time T_i,

$$F(t|\boldsymbol{x}_i) = 1 - \exp\{-\Lambda_0(t)\exp(\boldsymbol{x}_{i1}\boldsymbol{\beta}_1 + \boldsymbol{x}_{i2}\boldsymbol{\beta}_2)\}, \quad (7)$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2})'$, $\Lambda_0(t) = \log(1 + t) + t^{1/2}$, $x_{i1} \sim$ Bernoulli(0.5), and $x_{i2} \sim \text{N}(0, 0.5^2)$, for $i = 1, \ldots, n$. The sample size was specified to be $n = 200$ and all possible combinations of $\beta_1 \in \{-1, 0, 1\}$ and $\beta_2 \in \{-1, 0, 1\}$ were considered, resulting in 9 parameter configurations. Each T_i was generated by solving $F(t|\boldsymbol{x}_i) = u_i$ numerically, where $u_i \sim \mathcal{U}_{(0,1)}$. The number of observation times for each subject was generated according to 1 plus a Poisson random variable having mean parameter 6. Proceeding in this fashion ensured that each subject has at least one observation time, but allowed the number of observation times to vary from subject to subject. The gap times between adjacent observations were sampled according to an exponential distribution with mean 0.5. Subsequently, the observation times were given by the cumulative sums of the gap times. The observed interval for the ith subject was then determined to be the two consecutive observation times whose interval contained T_i, with the convention that if T_i was less (greater) than the smallest (largest) observation time then the lower (upper) bound of the observed interval was 0 (∞). For the purposes of this study, 500 data sets of the form $\{(L_i, R_i], \boldsymbol{x}_i\}_{i=1}^n$ were generated for each considered parameter configuration. The average rate of right-censoring varied from 3% to 21% across all configurations; see Table 1.

The cumulative baseline hazard function was modeled using basis splines having degree 3 and a knot set having cardinality 5 on the interval (t_{min}, t_{max}), where t_{min} and t_{max} are the minimum and maximum values of the set of observed interval end points excluding 0 and ∞. The interior knots were placed at the first, second, and third quartiles of the set of interval end points falling between t_{min} and t_{max}. A similar simulation study (results not shown) considering equally spaced knots over the interval (t_{min}, t_{max}) was also performed,

and it resulted in the same conclusions presented herein. The initial values for the EM algorithm were specified to be $\boldsymbol{\theta}^{(0)}=(\boldsymbol{\beta}^{(0)'},\gamma^{(0)'})'=(\mathbf{0}_2',\mathbf{1}_6')'$, although numerous additional simulation studies (results not shown) demonstrated that the algorithm is relatively robust, in terms of the accuracy of parameter estimation and convergence rate, to the choice of $\boldsymbol{\theta}^{(0)}$. Convergence of the EM algorithm was declared when the maximum absolute change between successive estimates of $\boldsymbol{\theta}$ was less than 0.005; i.e., $\max_h|\theta_h^{(d+1)}-\theta_h^{(d)}|<0.005$, where $\theta_h^{(d)}$ is the hth element of $\boldsymbol{\theta}^{(d)}$.

For purposes of comparisons, two competing techniques were implemented. The first technique fits the PH model via the ICM-algorithm (ICM) of Pan (1999), and was implemented using the `intcox` function in R (Henschel and Mansmann, 2013). The second technique, proposed by Zhang et al. (2010), makes use of a spline-based sieve semiparametric maximum likelihood (SML) approach to fit the PH model to interval-censored data, with $\Lambda_0(\cdot)$ being approximated through the use of monotone B-splines. These comparisons were chosen for a variety of reasons. In particular, ICM constitutes the only frequentist based approach that has a companion statistical package specifically designed for analyzing interval-censored data under the semiparametric PH model, while SML is the most recent contribution to the literature that is directly comparable with the proposed methodology. For each modeling technique, Table 1 summarizes the empirical bias and sample standard deviation of the 500 point estimates, the average of the 500 estimated standard errors, and the empirical coverage probability associated with 95% Wald confidence intervals for each of the regression parameters, as well as the average model fitting times.

From the results presented in Table 1, first note that the regression estimates obtained by the proposed method are all close to their corresponding true parameter values. Secondly, the sample standard deviation and the averaged standard errors of the 500 estimates are in agreement, indicating that the asymptotic approximation of the variance-covariance matrix obtained from Louis's method performs well for finite samples. Lastly, the empirical coverage probabilities for the confidence intervals for the regression parameters are predominantly at their nominal level, suggesting that the use of Wald-type inference may be appropriate for evaluating estimates obtained by the EM algorithm.

Comparing the proposed methodology to the two competing techniques, one will note that both in terms of parameter estimation and inferential characteristics the proposed methodology performed as well, if not better, than SML, across all considered configurations. In contrast, ICM yielded biased point estimates and does not provide estimated standard errors, as was pointed out in Section 1. These findings are congruous with the results presented in Pan (1999) regarding the performance of ICM. Though similar in terms of estimation and inference, the discernible advantage of the proposed methodology over that of SML arises in the model fitting times; i.e., SML took on average 10 to 25 times longer to complete model fitting when compared to the proposed methodology. This advantage could render the proposed approach preferable when analyzing larger data sets, as the model fitting times for both the proposed EM algorithm and SML increase with the sample size.

3.2 Simulation study II

The following simulation study assesses the performance of the proposed methodology under high right-censoring rates. In this study the failure time model in (7) was again considered with $\Lambda_0(t) = t/10 - \log(1 + t/10)$, $x_{i1} \sim$ Bernoulli(0.5), and $x_{i2} \sim N(0, 0.25^2)$, for $i = 1, \ldots, n$. The observational process described in Section 3.1 was again used, with the number of observation times being determined by 1 plus a Poisson random variable having mean parameter 1, and the gap times between adjacent observations were sampled according to an exponential distribution with mean 4. For each parameter configuration, 500 data sets were generated with each containing n observations, where $n \in \{200, 2000\}$. The average right-censoring rate varied from 71% to 85% across the 9 parameter configurations; see Table 2.

Table 2 summarizes the estimates of β obtained by the EM algorithm and the two competing methods, as well as the average model fitting times when $n = 200$. Web Table 1 provides the corresponding results when $n = 2000$. This summary again illustrates that the proposed technique performs well; i.e., the EM algorithm obtains estimates that exhibit little if any evidence of bias, results in accurate variance estimates, and produces confidence intervals that attain their nominal coverage probability. In contrast, SML encounters numerical instabilities which results in the algorithm terminating due to numerical error for a significant number (approximately 5%-10%) of the considered data sets, and this feature persists for larger values of n; see Web Table 1. For the data sets for which numerical instability was not encountered, SML continues to provide accurate estimates and reliable inference. In terms of computational burden the proposed method is again superior when compared to SML, in this setting. The estimates obtained from ICM again exhibit considerable bias.

3.3 Simulation summary

The results of the simulation studies presented in Section 3.1 and 3.2 demonstrate that the proposed methodology can be used to efficiently, accurately, and reliably analyze interval-censored data across a broad spectrum of censoring rates. The same cannot always be said for the two competing procedures. In addition to the simulation results presented herein, a summary of the estimation of the baseline cumulative distribution function F_0, and consequently the estimation of the cumulative baseline hazard function, is provided in Web Table 2, across all considered simulation configurations. Briefly, these findings indicate that the proposed method provides precise estimates of F_0 that are comparable to the estimates obtained by SML, and are superior to those resulting from ICM. Further, Web Appendix D provides two additional simulation studies: one that considers medium right-censoring rates and the other compares the proposed approach and SML in terms of model fitting times for larger sample sizes. The results from the former study reinforce the main findings discussed in Section 3.1, while the results of the latter study indicate that SML is far more computationally burdensome when compared to the proposed method; e.g., for $n = 50000$ observations the proposed approach took approximately 1 minute, on average, to complete model fitting which was more than 140 times faster than SML; see Web Table 3.

4. Data application

Sponsored by the United States National Cancer Institute, the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial was initiated in 1993 in an effort to assess the effect of routine screening on cancer-related mortality as well as other relevant secondary endpoints. Participants in this population-based randomized trial consisted of men and women between the ages of 55 and 74, who had no previous history of any PLCO cancer, and were not participating in any other cancer screening and/or primary prevention trials. At the time of enrollment, participants were randomized into either the control or intervention arm. Participants in the intervention arm received regular screenings for PLCO cancers during the first 6 years, and were followed for an additional 7 years. In contrast, participants randomized to the control arm were simply followed for 13 years after enrollment. For further details about the PLCO Cancer Screening Trial see Andriole et al. (2012). The data collected from this study consisted of screening results and various risk factors, e.g., age, race, etc.

This analysis considers the prostate cancer screening data collected on male participants in the intervention arm. In particular, this data consists of screening and follow up information which spans a 10 year period of time. During the first 6 years of this period, participants in the intervention arm were screened approximately once a year via a Prostate Specific Antigen (PSA) test. If abnormally high PSA levels were detected, indicating the possible development of prostate cancer, a prostate biopsy was performed to determine whether or not the participant had developed prostate cancer.

The primary focus of this analysis is to assess the association of risk factors with the time from randomization until the onset of prostate cancer. Due to the design of the study and the nature of prostate cancer, the onset times were not observable but rather were known relative to the screening times; i.e., they were interval-censored. In particular, of the 32720 observations having complete covariate information, 7 (0.02%) were left-censored, 2853 (8.7%) were interval-censored and 29860 (91.3%) were right-censored. In total, 12 covariates were considered: age (centered) at randomization; education, with 1 indicating a college education; race, with categories Caucasian, African American, and other; obesity, with 1 indicating obesity; heart, with 1 indicating presence of heart disease; stroke, with 1 indicating a previous stroke; diabetes, with 1 indicating diabetic; colitis, with 1 indicating a positive status; hepatitis, with 1 indicating a positive status; aspirin, with 1 indicating regular use; ibuprofen, with 1 indicating regular use; family history, with 1 indicating that an immediate relative had prostate cancer. For a summary of these risk factors see Web Table 4.

To analyze these data using the proposed methods, the cumulative baseline hazard function was modeled using basis splines having degree 3 and a candidate knot set consisting of m = 50 interior knots, which were equally spaced over the time domain, was considered. A backward elimination procedure based on AIC (BIC) was used to identify the final model which made use of m = 28 (19) interior knots; for a summary and discussion of the model fits based on this procedure see Web Appendix E.

The estimated regression coefficients obtained by the EM algorithm are summarized in Table 3 for the two final candidate models. For comparative purposes, the analysis was also attempted using SML. In each of the attempted implementations, the SML model fitting algorithm either terminated, due to numerical instabilities, or converged to a local extrema depending on the parameter initialization; see Web Appendix E for further discussion. In contrast, across all considered initializations and interior knot specifications, the proposed procedure resulted in practically identical estimates of the regression coefficients and inferential conclusions.

The proposed approach identified that race, family history, diabetes, and age were significant risk factors associated with the development of prostate cancer, while all other considered risk factors were insignificant. In particular, African American, family history, and age were found to be positively associated with the hazard of developing prostate cancer, while all other significant factors were negatively associated with the hazard of developing prostate cancer. Figure 1 provides a plot of the estimated survival function from the EM algorithm, when $m = 28$, at the different levels of race. Also included are the corresponding nonparametric estimates of the survival functions which were obtained according to the approach of Turnbull (1976). Web Figure 1 provides the analogous results for $m = 19$. From these figures, it appears that the PH model provides a good fit to these data.

5. Discussion

This paper proposes a new method for analyzing general interval-censored data under the proportional hazards model. Under a flexible parametric formulation of the PH model, an EM algorithm was developed that can be used to find the maximum likelihood estimates of all unknown parameters. The key step in deriving the algorithm involves expanding the observed data likelihood to a complete data likelihood through a two stage data augmentation procedure. This is achieved by linking the failure time under the PH model with a latent non-homogeneous Poisson process. The proposed EM algorithm is straightforward to implement, enjoys quick convergence, and provides simple closed-form variance estimates. A companion R package `ICsurv` has been developed and is publicly available from the CRAN (i.e., http://cran.us.r-project.org/); for further details see Web Appendix F. In summary, the proposed method provides an accurate, reliable, and computationally efficient approach that can be used to analyze interval-censored data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the Editor, the Associate Editor, and the two referees for their helpful suggestions. The authors thank Dr. Ying Zhang for providing the code used to implement SML. Michael G. Hudgens was partially supported by NIH grant R01 AI029168.

References

Allison, P. Survival analysis using SAS: A practical guide. 2nd. SAS Publishing; Cary, NC: 2010.

Andriole G, Crawford E, Grubb R, Buys S, Chia D, Church T, et al. Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. Journal of the National Cancer Institute. 2012; 104:125–132. [PubMed: 22228146]

Betensky R, Lindsey J, Ryan L, Wand M. A local likelihood proportional hazards model for interval-censored data. Statistics in Medicine. 2002; 21:263–275. [PubMed: 11782064]

Cai B, Lin X, Wang L. Bayesian proportional hazards model for current status data with monotone splines. Computational Statistics and Data Analysis. 2011; 55:2644–2651.

Cai T, Betensky R. Hazard regression for interval-censored data with penalized spline. Biometrics. 2003; 59:570–579. [PubMed: 14601758]

Cox R. Regression models and life-tables. Journal of the Royal Statistical Society: Series B. 1972; 4:187–220.

Finkelstein D. A proportional hazards model for interval-censored failure time data. Biometrics. 1986; 42:845–854. [PubMed: 3814726]

Goeteghebeur E, Ryan L. Semiparametric regression analysis of interval-censored data. Biometrics. 2000; 56:1139–1144. [PubMed: 11129472]

Goggins W, Finkelstein D, Schoenfeld D, Zaslavsky M. A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. Biometrics. 1998; 54:1498–1507. [PubMed: 9883548]

Gómez G, Calle M, Oller R, Langohr K. Tutorial on methods for interval-censored data and their implementation in R. Statistical Modeling. 2009; 9:259–297.

Groeneboom, P.; Wellner, J. Information Bounds and Non-Parametric Maximum Likelihood Estimation. Birkhauser; Boston: 1992.

Henschel, V.; Mansmann, U. intcox: Iterated convex minorant algorithm for interval-censored event data. R package version 0.9.3. 2013. http://CRAN.R-project.org/package=intcox

Li, J.; Ma, S. Chapman & Hall/CRC Biostatistic Series. CRC Press LLC; 2013. Survival Analysis in Medicine and Genetics.

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]

Liu H, Shen Y. A semiparametric regression cure model for interval-censored data. Journal of the American Statistical Association. 2009; 104:1168–1178. [PubMed: 20354594]

Louis T. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B. 1982; 44:226–233.

McMahan C, Wang L, Tebbs J. Regression analysis for current status data using the EM algorithm. Statistics in Medicine. 2013; 32:4452–4466. [PubMed: 23761135]

Odell P, Anderson K, D'Agostino R. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. Biometrics. 1992; 48:951–959. [PubMed: 1420849]

Pan W. Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. Journal of Computational and Graphical Statistics. 1999; 8:109–120.

Pan W. A multiple imputation approach to Cox regression with interval-censored data. Biometrics. 2000; 56:199–203. [PubMed: 10783796]

Ramsay J. Monotone regression splines in action. Statistical Science. 1988; 3:425–441.

Rosen J. The gradient projection method for nonlinear programming. Journal of the Society for Industrial and Applied Mathematics. 1960; 8:181–217.

Rosenberg P. Hazard function estimation using B-splines. Biometrics. 1995; 51:874–887. [PubMed: 7548706]

Rucker G, Messerer D. Remission duration: an example of interval-censored observations. Statistics in Medicine. 1988; 7:1139–1145. [PubMed: 3201039]

Satten G. Rank based inference in the proportional hazards model for interval-censored data. Biometrika. 1996; 83:355–370.

Shao F, Li J, Ma S, Lee M. Semiparametric varying-coefficient model for interval-censored data with a cured proportion. Statistics in Medicine. 2014; 33:1700–1712. [PubMed: 24302535]

Sun, J. The Statistical Analysis of Interval-Censored Data. Springer; Berlin: 2006.

Turnbull B. The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B. 1976; 38:290–295.

Wang L, Dunson D. Semiparametric Bayes proportional odds models for current status data with under-reporting. Biometrics. 2011; 67:1111–1118. [PubMed: 21175554]

Wang L, Lin X. A Bayesian approach for analyzing case 2 interval-censored failure time data under the semiparametric proportional odds model. Statistics and Probability Letters. 2011; 81:876–883.

Zhang Y, Hua L, Huang J. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. Scandinavian Journal of Statistics. 2010; 37:338–354.

Zhang Y, Jamshidian M. On algorithms for the nonparametric maximum likelihood estimator of the failure function with censored data. Journal of Computational and Graphical Statistics. 2004; 13:123–140.

Zhang Z, Sun J. Interval-censoring. Statistical Methods in Medical Research. 2010; 19:53–70. [PubMed: 19654168]

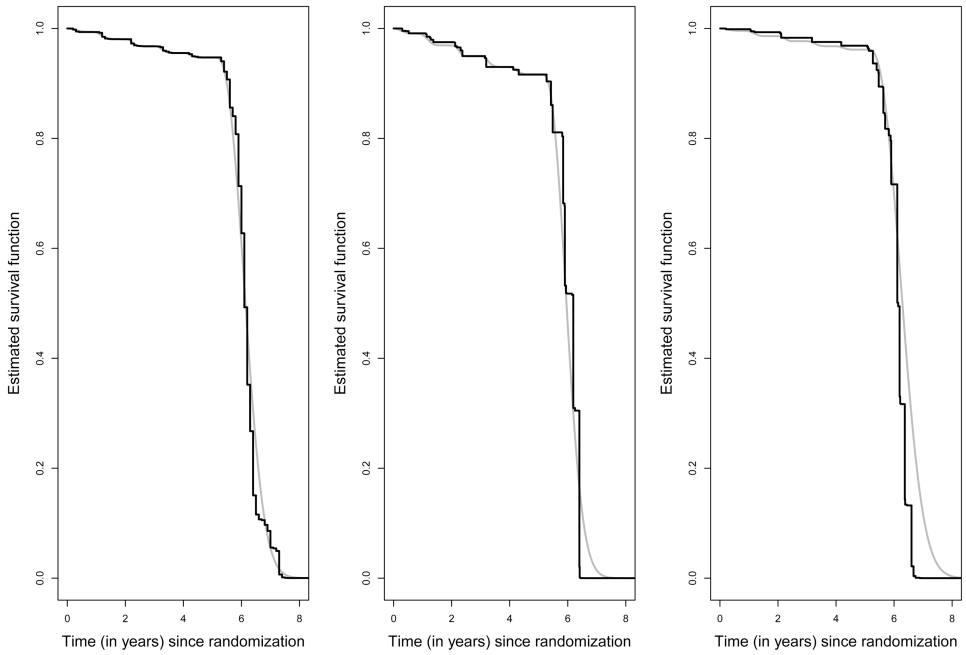

Figure 1.

PLCO data analysis: Estimates of the survival functions obtained by the proposed method (smooth gray curves) and the Turnbull estimates (black step functions) at the different levels of race: Caucasian (left panel), African American (center panel), and other (right panel). These estimates were obtained by first dividing the 32720 observations into three strata based on race. The Turnbull estimates were obtained within each of these strata separately. For the PH model, a survival curve was estimated for each observation and these estimates were then averaged within each strata providing the depicted estimated survival curve for the PH model.

Table 1

Simulation study I: Empirical bias (Bias) and standard deviation (SD) of the 500 estimates of β, the average of the estimated standard errors (ESE), and the empirical coverage probabilities associated with 95% Wald confidence intervals (CP95). Also provided are the right-censoring rates (RR) under the different simulation settings and the average model fitting time (Time) in seconds.

	EM					SML					ICM			
Parameter	Bias	SD	ESE	CP95	Time	Bias	SD	ESE	CP95	Time	Bias	SD	Time	RR
$\beta_1 = -1$	-0.02	0.18	0.18	0.96	0.77	-0.02	0.18	0.19	0.96	15.80	0.31	0.18	0.27	20.1%
$\beta_2 = -1$	-0.03	0.20	0.19	0.93		-0.03	0.20	0.20	0.93		0.06	0.18		
$\beta_1 = 0$	0.00	0.16	0.16	0.95	0.90	0.00	0.16	0.18	0.97	15.87	0.23	0.17	0.30	8.5%
$\beta_2 = -1$	-0.05	0.19	0.19	0.96		-0.05	0.19	0.21	0.97		-0.01	0.18		
$\beta_1 = 1$	0.03	0.19	0.20	0.95	1.22	0.04	0.21	0.23	0.98	13.16	0.19	0.21	0.38	4.7%
$\beta_2 = -1$	-0.04	0.21	0.21	0.95		-0.04	0.22	0.24	0.97		-0.01	0.21		
$\beta_1 = -1$	-0.02	0.18	0.17	0.95	0.76	-0.02	0.18	0.18	0.96	17.49	0.30	0.18	0.23	19.2%
$\beta_2 = 0$	0.00	0.18	0.17	0.93		0.00	0.18	0.18	0.94		0.01	0.17		
$\beta_1 = 0$	-0.01	0.17	0.16	0.94	0.71	-0.01	0.17	0.17	0.95	19.07	0.22	0.18	0.27	6.0%
$\beta_2 = 0$	-0.02	0.16	0.16	0.94		-0.02	0.16	0.18	0.96		-0.01	0.16		
$\beta_1 = 1$	0.04	0.20	0.20	0.96	1.55	0.05	0.20	0.23	0.97	17.64	0.20	0.22	0.33	3.1%
$\beta_2 = 0$	0.00	0.17	0.18	0.96		0.00	0.17	0.21	0.98		0.00	0.17		
$\beta_1 = -1$	-0.02	0.17	0.18	0.96	0.77	-0.02	0.17	0.19	0.97	15.07	0.31	0.18	0.27	21.3%
$\beta_2 = 1$	0.04	0.21	0.19	0.93		0.04	0.21	0.20	0.95		-0.05	0.19		
$\beta_1 = 0$	0.01	0.16	0.16	0.94	0.90	0.01	0.16	0.18	0.96	14.74	0.24	0.17	0.30	8.6%
$\beta_2 = 1$	0.04	0.21	0.19	0.94		0.04	0.21	0.21	0.95		-0.01	0.19		
$\beta_1 = 1$	0.03	0.19	0.20	0.96	1.15	0.03	0.20	0.23	0.97	12.40	0.19	0.20	0.36	4.7%
$\beta_2 = 1$	0.04	0.21	0.21	0.96		0.04	0.21	0.24	0.97		0.01	0.20		

Table 2

Simulation study II for n = 200: Empirical bias (Bias) and standard deviation (SD) of the 500 estimates of β, the average of the estimated standard errors (ESE), and the empirical coverage probabilities associated with 95% Wald confidence intervals (CP95). Also provided are the right-censoring rates (RR) under the different simulation settings and the average model fitting time (Time) in seconds. Further, the percentage of the data sets for which SLM failed to converge are reported, in parenthesis, along with this procedure's average model fitting time.

	EM					SML					ICM			
Parameter	Bias	SD	ESE	CP95	Time	Bias	SD	ESE	CP95	Time	Bias	SD	Time	RR
$\beta_1 = -1$	-0.04	0.44	0.51	0.95	0.83	-0.04	0.45	0.49	0.97	30.07(9.4%)	0.66	0.29	0.76	85.2%
$\beta_2 = -1$	0.01	0.87	0.83	0.92		0.00	0.87	0.91	0.97		0.36	0.57		
$\beta_1 = 0$	0.02	0.35	0.44	0.96	0.92	0.01	0.35	0.36	0.98	27.77(7.6%)	0.33	0.30	0.58	79.8%
$\beta_2 = -1$	-0.07	0.71	0.69	0.94		-0.07	0.71	0.75	0.97		0.16	0.53		
$\beta_1 = 1$	0.05	0.32	0.40	0.93	1.14	0.05	0.32	0.33	0.96	29.21(3.8%)	0.15	0.24	0.54	71.2%
$\beta_2 = -1$	-0.01	0.64	0.59	0.92		0.00	0.64	0.63	0.95		0.19	0.52		
$\beta_1 = -1$	-0.06	0.46	0.59	0.96	0.78	-0.08	0.47	0.50	0.98	28.22(9.8%)	0.65	0.27	0.74	85.6%
$\beta_2 = 0$	0.04	0.84	0.81	0.93		0.02	0.84	0.92	0.97		0.03	0.52		
$\beta_1 = 0$	0.01	0.32	0.46	0.96	0.73	0.01	0.33	0.36	0.98	29.27(9.4%)	0.32	0.28	0.57	80.0%
$\beta_2 = 0$	-0.03	0.67	0.67	0.95		-0.02	0.68	0.74	0.97		-0.01	0.56		
$\beta_1 = 1$	0.04	0.31	0.41	0.95	1.06	0.04	0.31	0.32	0.97	28.91(2.2%)	0.15	0.23	0.52	70.9%
$\beta_2 = 0$	0.02	0.58	0.57	0.95		0.02	0.59	0.61	0.96		0.02	0.49		
$\beta_1 = -1$	-0.06	0.46	0.56	0.95	0.87	-0.05	0.46	0.49	0.99	28.35(9.4%)	0.65	0.29	0.77	85.3%
$\beta_2 = 1$	0.06	0.79	0.85	0.94		0.05	0.79	0.92	0.97		-0.34	0.50		
$\beta_1 = 0$	0.00	0.34	0.45	0.95	0.99	0.00	0.34	0.36	0.97	29.22(8.2%)	0.31	0.29	0.62	79.8%
$\beta_2 = 1$	0.04	0.72	0.69	0.92		0.02	0.71	0.75	0.95		-0.17	0.55		
$\beta_1 = 1$	0.05	0.33	0.38	0.92	1.32	0.04	0.33	0.33	0.96	29.97(4.4%)	0.16	0.23	0.54	71.2%
$\beta_2 = 1$	0.02	0.61	0.58	0.94		0.01	0.61	0.63	0.96		-0.17	0.50		

Table 3

PLCO data analysis: Estimated regression coefficients for the covariates, estimated standard errors (ESE), and p-values obtained by the proposed approach. Presented results are from the two final models that were selected by BIG and AIC which use m=19 and 28 interior knots to model the cumulative baseline hazard function, respectively.

	$m = 19$			$m = 28$		
Covariate	**Estimate**	**ESE**	**P-value**	**Estimate**	**ESE**	**P-value**
Race(African American)	0.528	0.098	0.000	0.529	0.101	0.000
Race (Other)	-0.307	0.112	0.006	-0.320	0.115	0.005
Education	0.015	0.059	0.792	0.016	0.075	0.836
Obesity	-0.072	0.059	0.220	-0.077	0.063	0.219
Heart	-0.059	0.072	0.413	-0.060	0.072	0.404
Stroke	-0.153	0.155	0.322	-0.162	0.154	0.293
Diabetes	-0.456	0.097	0.000	-0.452	0.097	0.000
Colitis	-0.057	0.228	0.804	-0.061	0.230	0.792
Hepatitis	-0.078	0.129	0.547	-0.084	0.129	0.514
Aspirin	-0.014	0.052	0.787	-0.014	0.061	0.812
Ibuprofen	0.030	0.055	0.588	0.031	0.058	0.597
Family history	0.444	0.074	0.000	0.454	0.077	0.000
Age	0.056	0.004	0.000	0.057	0.005	0.000