Biostatistics (2013), **14**, 4, *pp.* 766–778
doi:10.1093/biostatistics/kxt019
Advance Access publication on May 30, 2013

Stochastic EM algorithm for doubly interval-censored data

DAVID DEJARDIN*

Interuniversity Institute for Biostatistics and Statistical Bioinformatics, KULeuven and Universiteit Hasselt, Kapucijnenvoer 35, Blok D, Bus 7001, B3000 Leuven, Belgium

Global Biometric Sciences, Bristol-Myers Squibb, 1420 Braine l'Alleud, Belgium

david.dejardin@med.kuleuven.be

EMMANUEL LESAFFRE

Department of Biostatistics, Erasmus MC, 3015 CE Rotterdam, The Netherlands

Interuniversity Institute for Biostatistics and Statistical Bioinformatics, KULeuven and Universiteit Hasselt, Kapucijnenvoer 35, Blok D, Bus 7001, B3000 Leuven, Belgium

Summary

In clinical trials, it is frequently of interest to estimate the time between the onset of two events (e.g. duration of response in oncology). Here, we consider the case where subjects are assessed at fixed visits but the initial event and the terminating event occur in between visits. This type of data, called doubly interval censored, is often analyzed with standard survival techniques, assuming either that the survival time (between initial and terminating event) is known exactly or is single interval censored. We introduce a motivating dataset in which the interest is to evaluate the impact of the treatment on the duration of response endpoint. We review the existing approaches and discuss their limitations with respect to the characteristics of our motivating dataset. Furthermore, we propose a stochastic EM algorithm that overcomes the problems in the existing approaches. We show by simulations the finite sample properties of our approach.

Keywords: Cox proportional hazard; Doubly interval censored; Stochastic EM algorithm.

1. Introduction

In most survival studies, the time of origin of the survival time is known or assumed to be known. However, it also occurs that the start of the period at risk is only known to lie in an interval. In clinical trials, the origin of the survival is often the time of randomization (and the start of treatment) but it could also be the time that a patient enters a particular state. An example of the latter case occurs in HIV research where the onset of HIV can only be established at a doctor's visit. The state can only be known to happen between visits and then the time to HIV infection is called interval censored. In addition, since most often

*To whom correspondence should be addressed.

the end of the period at risk is right censored or interval censored, we obtain doubly right (DR)-censored or doubly interval (DI)-censored survival times. Dedicated survival techniques are required to analyze this kind of data.

In this paper, we focus on techniques that evaluate the impact of covariates on the survival distribution of the DI survival time through a semi-parametric Cox proportional hazard (PH) model. This paper is motivated by the analysis of duration of response in a clinical trial in first-line metastatic breast cancer. Current methods are reviewed in the context of our motivating example. This dataset exhibits the following complexities. Firstly, both the time to initial event (response) and the time to terminating event (progression) are interval censored. Secondly, the discretization of the data (i.e. to assign the boundaries of the interval to a grid of values that is fixed and common to all subjects) is not a good option because of the irregular interval width (depending on whether the progression happens while taking the drug or during the follow-up period). We shall also discuss the ability of the existing methods to handle a data characteristic coined as "overlapping" that occurs when the upper bound of the observed interval of U exceeds the lower bound of the observed interval of V. This characteristic is not present in our motivating dataset but occurs in other applications we encountered. More generally, we shall review the existing methods in the context of controlled clinical trials (as our motivating example) for which the methods are to be specified prior to the analysis and should not depend on data-driven quantities (such as optimal bandwidth, a smoothing parameter, or predefined mass points). These methods might be problematic since the quantity may not be fixed in advance and different choices will lead to different results. Further, in the context of clinical trials, Bayesian methods may not be preferred either if they rely on fully parametric models or need informative priors. Therefore, we propose in this paper a semi-parametric approach that we believe is particularly suitable for the analysis of controlled clinical trials. More generally, we argue that our approach is particularly appealing when statistical methods need to be fully described in advance.

We use the following notation: Let U be the time to initial event at which the subject starts to be at risk for a condition and V be the time to terminating event at which the subject experiences the condition.

Our interest is in estimating the impact of covariates on the distribution of the survival time $T = V - U$, denoted $S_T(t|X) = 1 - F_T(t|X) = S_{0,T}(t)^{\exp(\boldsymbol{\beta}_T^{\mathrm{T}} X)}$. We refer to $S_{0,T}(t) = S_T(t|X = 0)$ as the baseline distribution, and $\lambda_0(t)$ denotes the corresponding baseline hazard with $\Lambda_0(t)$ the cumulative hazard ($\Lambda_0(t) = \int_0^t \lambda_0(x)\mathrm{d}x$). The density of T is denoted by f_T. The covariates are denoted as X and their parameters as $\boldsymbol{\beta}_T$, abbreviated as $\boldsymbol{\beta}$. Similarly, $S_U(u|X) = 1 - F_U(u|X)$ denotes the distribution of U and f_U its density and depends on covariates in a PH manner.

Both U and V are assumed to be interval censored, the observed data consist of an upper and lower bound for both variables denoted by $[U_l, U_r]$ for U and $[V_l, V_r]$ for V with realizations of these intervals denoted by $[u_{li}, u_{ri}]$ and $[v_{li}, v_{ri}]$, respectively, and the corresponding covariates X_i for subject i ($i = 1, \ldots, n$). In datasets where no overlapping occurs (as in our motivating example), we have the following relationship: $u_{li} \leqslant u_{ri} \leqslant v_{li} \leqslant v_{ri}$ for $i = 1, \ldots, n$. Overlapping occurs when $u_{ri} > v_{li}$. While our focus is on the estimation of $\boldsymbol{\beta}$ in the presence of DI survival times, right-censored V data (DR survival times) are allowed by setting the upper bound of the interval for V to ∞. Further, we assume here independent censoring (i.e. censoring that implies that both U and T are independent of the monitoring times from which U_l, U_r, V_l, and V_r are obtained; see Oller *and others* (2004) for details) and independence between U and T, which are classical assumptions for the treatment of DI survival times.

We first review existing methods and describe their limitations with respect to the characteristics of our dataset detailed above. In Section 3, we describe a novel approach to estimate the impact of a covariate in the presence of DI data. Section 4 is devoted to a limited simulation study evaluating our method in comparison with some existing methods. The analysis on the motivating dataset is given in Section 5. We end the paper with a discussion.

2. EXISTING APPROACHES

In the presence of DI or DR survival times, one may wish to focus only on the distribution of T, and ignore the distribution of U. Doing so leads to single interval-censored or right-censored data that can be analyzed with standard statistical methods and software. For DR data, Law and Brookmeyer (1992) have investigated the performance of the midpoint imputation (see Appendix A of supplementary material available at *Biostatistics* online for definition) for DR data in different contexts (estimation of the distribution, estimation of the hazard ratio). They showed that the bias increases and the coverage probability of the confidence interval decreases as the width of the interval of U increases. In Appendix A of supplementary material available at *Biostatistics* online, we define the different approaches that ignore the distribution of U and extend the simulations from Law and Brookmeyer (1992) to DI data. Therein, we also show the impact of misspecifying the distribution of U using a fully parametric approach.

De Gruttola and Lagakos (1989) have introduced a method for estimating the distribution in the presence of DI data. Their method is based on discretizing the distribution based on prespecified mass points. Kim *and others* (1993) used the same idea to propose an extension of the Cox PH model to DI data. These authors pointed out that identifiability problems may occur, especially when two points are jointly included or excluded from all intervals or when some observed intervals do not contain any mass points. So, it is difficult to specify the mass points in advance without reference to the actual observed data.

Sun *and others* (1999, 2004) proposed to estimate the regression coefficients in F_T by integrating out U (utilizing the fact that F_U can be estimated from the data, e.g. using the Turnbull estimator (Turnbull, 1976)). Goggins *and others* (1999) described a Monte-Carlo EM algorithm for a PH model for T. Pan (2001) proposed a multiple imputation approach. However, these approaches are restricted to DR data and are therefore not suitable for our motivating dataset.

Flexible Bayesian parametric approaches have been proposed by Komárek and Lesaffre (2008) and Jara *and others* (2010). Both approaches allow dependence between U and T, but have to a large extent a parametric nature for which non-informative priors may be difficult to derive formally. Therefore, we argue that these two methods are likely to be more suitable in an explorative analysis rather than to be used in a formal statistical analysis of a randomized clinical trial.

We now propose a novel computational approach to analyze semi-parametrically the impact of covariates on the distribution of T in the presence of DI survival times (i.e. it allows interval-censored V). Our method does not rely on prespecified mass points and accounts for the impact of covariates on U.

3. STOCHASTIC EM ALGORITHM TO ESTIMATE THE DISTRIBUTION OF T

3.1 *Concept*

The likelihood for DI data can be written as

$$L(\psi, \theta | u_i \in [u_{li}, u_{ri}], v_i \in [v_{li}, v_{ri}],\ i = 1, \ldots, n) = \prod_{i=1}^{n} \int_{u_{li}}^{u_{ri}} \int_{v_{li}}^{v_{ri}} f_U(u|\psi, X) f_T(v - u|\theta, X)\, \mathrm{d}u\, \mathrm{d}v, \tag{3.1}$$

where ψ and θ are the parameters of the unknown densities f_U and f_T. We are interested in the estimation of θ with minimal assumptions (ψ are treated as nuisance parameters and are omitted in the notation in the remainder of the paper). Without assumptions on f_T, the integral in (3.1) cannot be computed.

When U is observed and T right censored, the semi-parametric full likelihood underlying the Cox PH model can be written as (see Klein and Moeschberger, 2003)

$$L(\theta|y_i, \delta_i,\ i = 1, \ldots, n) = \prod_{i=1}^{n} (\exp(\boldsymbol{\beta}^{\mathrm{T}} X_i)\lambda(y_i))^{\delta_i} \exp(-\Lambda(y_i)\exp(\boldsymbol{\beta}^{\mathrm{T}} X_i)), \tag{3.2}$$

where y_i, δ_i is the right-censored datum, and $\delta_i = 1$ is the event indicator ($v_{ri} < \infty$) and 0 otherwise. No assumptions on f_T are required when using the usual semi-parametric Cox PH model. Here $\theta = \{\boldsymbol{\beta}, \lambda_1, \ldots, \lambda_d\}$, where $\lambda_1, \ldots, \lambda_d$ are hazard parameters, d is the number of events and $\lambda(y_i) = \lambda_i$ when y_i is the event time and 0 otherwise, $\Lambda(y_i) = \sum_{j:y_j \leqslant y_i} \lambda_j$. The $\boldsymbol{\beta}$ parameters are estimated by the partial likelihood technique and the hazard parameters by the Breslow estimator. However, y_i is not exactly observed but is either interval censored (i.e. $y_i \in [v_{li} - u_i, v_{ri} - u_i]$) when $\delta_i = 1$ or right censored (i.e $y_i = v_{li} - u_i$) when $\delta_i = 0$, where u_i is also not exactly observed but lies in $[u_{li}, u_{ri}]$.

The proposed method consists in assuming that U and T are unobserved (but known to lie in observed intervals), and uses a missing data technique, namely, the EM algorithm Dempster *and others* (1977), to derive the parameters of interest based on the right-censored data likelihood.

In the EM algorithm, the E-step computes the expectation of the log likelihood with respect to the missing values (here u_i and y_j), given the observed data and the parameters at previous iteration. We denote the expected log likelihood as $Q_{k+1}(\theta|\theta^k)$ at iteration $k+1$ with θ^k the parameters at iteration k.

$$Q_{k+1}(\theta|\theta^k) = \mathrm{E}_{\substack{u_i, \forall i \in N \\ y_j, \forall j \in D}} [\log L(\theta|y_i, \delta_i, \forall i \in N) \\ |X_i, u_i \in [u_{li}, u_{ri}] \forall i \in N,\ y_j \in [v_{lj} - u_j, v_{rj} - u_j] \forall j \in D,\ y_s = v_{ls} - u_s \forall s \in C; \theta^k],$$

where $N = \{1, \ldots, n\}$, D is the subset of N representing the observations for which V is interval censored, and C is the subset of N for which V is right censored ($N = D \cup C$). For simplicity of the notation, we denote the observed data as

$$\mathcal{D} = \{u_i \in [u_{li}, u_{ri}] \forall i \in N,\ y_j \in [v_{lj} - u_j, v_{rj} - u_j] \forall j \in D,\ y_s = v_{ls} - u_s \forall s \in C\}.$$

From Appendix A, we see that $Q_{k+1}(\theta|\theta^k)$ is constructed from the following distributions:

$$F_U(u_i|X_i, u_i \in [u_{li}, u_{ri}], v_i \in [v_{li}, v_{ri}], \theta^k), \tag{3.3}$$

for $i \in N$ and, for $j \in D$ (when $\delta_j = 1$), $F_T(y_j|X_j, u_j, y_j \in [v_{lj} - u_j, v_{rj} - u_j], \theta^k)$.

Equation (3.3) represents the conditional distribution of U given the data (abbreviated as $F_U(u|\mathcal{D})$). We note that this expression depends on the observed interval of V and on θ^k. We now motivate this dependence by a trivial example. Suppose that U and T are discrete random variables such that only two values $\tilde{u}_1$ and $\tilde{u}_2$ of U can fall in the observed interval $[u_l, u_r]$. Also, suppose that F_T is degenerate and can take only one value $\tilde{t}$. We observe the interval $[v_l, v_r]$ for V such that $\tilde{u}_1 + \tilde{t} \in [v_l, v_r]$ and $\tilde{u}_2 + \tilde{t} \notin [v_l, v_r]$. The observed interval of V and the distribution of T allow in this simple case to exclude the value $\tilde{u}_2$ from the possible values of U given the data and give information on $F_U(u|\mathcal{D})$. Hence, $F_U(u|\mathcal{D})$ depends on the observed interval for V and θ^k. The derivation of $F_U(u|\mathcal{D})$ as a function of the marginal distribution of F_U and F_T is given in Appendix A. This derivation provides a more mathematical justification of the dependence of $F_U(u|\mathcal{D})$ on $[v_{li}, v_{ri}]$ and θ^k.

The M-step subsequently maximizes Q_{k+1} with respect to θ. However, no closed form for Q_{k+1} can be derived and thus we cannot maximize Q_{k+1} easily in the M-step. Therefore, we propose instead to use the stochastic EM (StEM) algorithm introduced by Celeux and Diebolt (1985). The details of the algorithm are given below.

3.2 *Implementation of the StEM approach*

We now describe the proposed StEM algorithm for DI-censored data. For simplicity of the notation, we will consider the case of a single covariate. Extension to multiple covariates is straightforward.

Initialization:

(1) Obtain the initial estimate $\hat{S}^0_{0,T}$ and $\hat{\beta}^0$ from midpoint imputation.

StE-step 1 (Stochastic E-step):

(2) Generate at iteration $k+1$:

$$\bar{u}^{k+1}_{1q}, \ldots, \bar{u}^{k+1}_{nq}, \quad q = 1, \ldots, m,$$

from

$$\hat{F}_U(u_i | X_i, u_i \in [u_{li}, u_{ri}], v_i \in [v_{li}, v_{ri}], (\hat{S}^k_{0,T})^{\exp(\hat{\beta}^k X_i)}), \quad i = 1, \ldots, n. \tag{3.4}$$

Appendix B describes how (3.4) can be obtained as a piecewise quadratic expression but monotone increasing that can therefore be easily inverted to generate $\bar{u}^{k+1}_{1q}, \ldots, \bar{u}^{k+1}_{nq}$.

StE-step 2:

(3) Generate at iteration $k+1$:

$$\bar{y}^{k+1}_{jq} \forall j \in D, \quad q = 1, \ldots, m,$$

from

$$\{\hat{S}^k_{0,T}(y_j \mid \bar{u}^{k+1}_{jq}, \ y_j \in [v_{lj} - \bar{u}^{k+1}_{jq}, v_{rj} - \bar{u}^{k+1}_{jq}])\}^{\exp(\hat{\beta}^k X_i)} \quad \forall j \in D, \tag{3.5}$$

where (3.5) is obtained by $(\hat{S}^k_{0,T}(v_{lj} - \bar{u}^{k+1}_{jq}) - \hat{S}^k_{0,T}(y_j))/(\hat{S}^k_{0,T}(v_{lj} - \bar{u}^{k+1}_{jq}) - \hat{S}^k_{0,T}(v_{rj} - \bar{u}^{k+1}_{jq}))$, with $\hat{S}^k_{0,T}$ estimated using the Breslow estimator. Since the resulting estimator is piecewise constant, the $\bar{y}^{k+1}_{jq}$ $j \in D$ are sampled from a finite set of values. Recall that, for $s \in C$, y_s are fixed to $y_s = v_{ls} - u_s$.

M-step:

(4) Compute

$$\hat{\Lambda}^{k+1}_{0,q}(t) \quad \text{and} \quad \hat{\beta}^{k+1}_q,$$

from $\bar{y}^{k+1}_{jq} \forall j \in D$ and $\bar{y}_{sq} = v_{ls} - \bar{u}^{k+1}_{sq} \forall s \in C$, $q = 1, \ldots, m$. Since $\bar{y}^{k+1}_{1q}, \ldots, \bar{y}^{k+1}_{nq}$ represent right-censored survival times, we use partial likelihood and the Breslow estimator of the baseline hazard to obtain $\hat{\Lambda}^{k+1}_{0,q}(t)$ and $\hat{\beta}^{k+1}_q$.

(5) The $(k+1)$th intermediate estimates will be

$$\hat{\beta}^{k+1} = 1/m \sum_{q=1}^{m} \hat{\beta}^{k+1}_q, \quad \hat{\Lambda}^{k+1}_0(t) = 1/m \sum_{q=1}^{m} \hat{\Lambda}^{k+1}_{0,q}(t), \quad \text{and} \quad \hat{S}^{k+1}_{0,T}(t) = \exp(\hat{\Lambda}^{k+1}_0(t)).$$

As shown in Nielsen (2000), the StEM algorithm leads to a Markov chain $\hat{\beta}^{k+1}_q$ $(q = 1, \ldots, m)$ and a Markov chain composed of $\hat{\Lambda}^{k+1}_{0,q}(t)$ $(q = 1, \ldots, m)$ representing the distribution of $\hat{\beta}$ and $\hat{\Lambda}_0$ at time t. Convergence is checked, e.g. by comparing the running mean of the StEM iterations to the current estimate through the mean integrated squared error. Further, Nielsen (2000) establishes the asymptotic normality of the StEM estimates at convergence (under some regularity conditions).

In Appendix C of supplementary material available at *Biostatistics* online, we show that the choice of m is not critical if larger than 50.

3.3 *Calculation of the variance of the parameters* $\boldsymbol{\beta}$

The variance of the estimated parameters obtained by the EM algorithm has to account for the sampling variability but also for the extra uncertainty due the missing data. Louis (1982) provided the theory to estimate this variance. His proposal is to derive an expression for the information matrix of the estimated parameters based on the observed data (denoted I_O), which is derived from the information matrix of the estimated parameters for the right-censored data I_T and the score vector of the right-censored data likelihood. Our estimator extends the estimator given by Goggins *and others* (1999) for DR data to DI data.

Louis' formula for I_O is described in the setting where the asymptotic maximum likelihood theory applies (in particular, it requires a finite number of parameters). In the particular case of the Cox PH model like we used in the StEM algorithm, the restriction to the finite number of parameters setting poses three practical issues: First, we evaluate the information matrix on both $\boldsymbol{\beta}$ and the parameters of $S_{0,T}(t)$. The number of parameters for $S_{0,T}(t)$ increases with the number of events. Second, the Louis's estimator assumes that the parameters are common to each of m datasets, which is not the case as the parameters for $S_{0,T}(t)$ pertain to "observed" death times, which are different for each randomly generated dataset. Thirdly, the fact that the values $\bar{y}_j$, $j \in D$ are generated from the piecewise constant Breslow estimator implies that ties can occur in the dataset. The presence of ties implies that the dimension of the information matrix is not constant across datasets.

However, to overcome the first and second issue (assuming for now a constant number of parameters), we note that the information matrix and score vector for $\hat{S}_{0,T}(t)$ are based on the risk set at each event time (i.e. on relative ordering of censored observations and events) and not on the location of the event times. Therefore, we treat the parameters for $S_{0,T}(t)$ as common to each generated dataset even though they are pertaining to different times. By doing so, and as we are interested only in the variance of $\hat{\boldsymbol{\beta}}$, we account for the impact of the parameters of $S_{0,T}(t)$ on the variance estimator of $\hat{\boldsymbol{\beta}}$. See also Appendix C for an elaboration of this argument.

To overcome the third issue, we note that it is possible to avoid ties by adding an EM iteration after the convergence of the algorithm based on a piecewise linear estimator of $S_{0,T}(t)$ (instead of piecewise constant). The details of the variance calculation are given in Appendix C.

4. Simulations

To assess the performance of the StEM algorithm, we have performed a limited simulation study. Datasets were generated as follows: Scenarios 1–5 investigated the setting in which the distribution of U is independent of covariates. Values of U were generated from an exp(1) distribution for Scenarios 1, 2, 4, and 5 and from a Weibull(2, 5) in Scenario 3. Values of T were generated from $S_{0,T}(t)^{\exp(\beta X)}$ where $S_{0,T}$ is a Weibull(1.7, 5.83) for Scenarios 1 and 2, a log-normal(2, 0.3) for Scenario 3 and a Weibull(2,5) for Scenarios 4 and 5. X was a binary covariate for Scenarios 1, 3, 4, and 5, and a uniform [0, 1] covariate for Scenario 2. Scenarios 6 and 7 investigated the setting in which U and T depend on common covariate(s). In Scenario 6, a single binary covariate was used, while in Scenario 7, X_1 was binary and X_2 was uniform [0, 1]. U was generated from $F_U(u|X) = 1 - \exp(-u)^{\exp(\boldsymbol{\beta}_U^{\mathrm{T}} X)}$ with $\beta_U = 0.5$ for Scenario 6 (binary covariate) and for Scenario 7, $\beta_U^1 = 0.5$ (binary) and $\beta_U^2 = -0.5$ (continuous). For both Scenarios $S_{0,T}$ is from a Weibull(2, 5). Intervals for U and $V = U + T$ were constructed by generating uniform random cutpoints for U and V. The number of cutpoints is given in Table 1. To ensure that the intervals cover most

Table 1. *Simulation results for the StEM algorithm with covariates*

Scenario	Nb cutpoints	Sample size	β	Estimator	$\hat{\beta}$	STE	STD	95% coverage probability	Power
Covariates on T only	1–6 cutpoints	200	0.5	Midpoint for U and V	0.36	0.15	0.15	0.83	0.71
				Reduced formulation	0.58	0.26	0.19	0.81	0.79
				StEM	0.53	0.21	0.20	0.95	0.76
	2–6 cutpoints X cont.	200	−0.5	Midpoint for U and V	−0.39	0.25	0.25	0.94	0.34
				Reduced formulation	−0.53	0.33	0.30	0.93	0.43
				StEM	−0.51	0.30	0.31	0.96	0.38
	3–15 cutpoints	100	0.5	Midpoint for U and V	0.40	0.22	0.21	0.92	0.52
				Reduced formulation	0.58	0.32	0.26	0.88	0.62
				StEM	0.51	0.27	0.26	0.95	0.52
	10 cutpoints	100	0.5	Midpoint for U and V	0.34	0.21	0.21	0.86	0.62
				Reduced formulation	0.81	0.60	0.30	0.75	0.71
				StEM	0.59	0.36	0.32	0.95	0.43
	4–6 cutpoints	100	0.5	Midpoint for U and V	0.34	0.22	0.21	0.86	0.39
				Reduced formulation	0.63	0.45	0.28	0.81	053
				StEM	0.56	0.36	0.34	0.94	0.37
	5–10 cutpoints	50	0.5	Midpoint for U and V	0.43	0.30	0.30	0.94	0.27
				Reduced formulation	0.60	0.48	0.37	0.9	0.36
				StEM	0.57	0.41	0.39	0.95	0.29
Single binary covariate impacting U and T	6–6 cutpoints	100	−0.5	Midpoint for U and V	−0.39	0.22	0.21	0.90	0.46
				Reduced formulation	−0.56	0.33	0.26	0.88	0.53
				StEM	−0.52	0.28	0.26	0.93	0.54
Two covariates impacting U and T	7–10 cutpoints	100	1	Midpoint for U and V	0.91	0.22	0.22	0.94	1
			−0.5		−0.47	0.36	0.36	0.96	0.24
			1	Reduced formulation	1.14	0.27	0.27	0.87	0.99
			−0.5		−0.56	0.46	0.41	0.97	0.25
			1	StEM	1.05	0.26	0.26	0.96	1
			−0.5		−0.53	0.38	0.42	0.96	0.20

Mean estimates of β are given along with the estimated STD, square root of the variance of the estimated parameter values (STE), 95% coverage probability, and power to detect a difference (Wald test with 0.05 significance level). X cont. means X continuous.

of the generated values, the cutpoints for U were generated within $[0, F_U^{-1}(0.99)]$ and the cutpoints for V were generated within $[0, F_U^{-1}(0.99) + F_T^{-1}(0.99)]$. Observations falling outside of this range were right censored.

The StEM approach (with $m = 100$) is compared with reduced likelihood methods: (1) the midpoint for U and V to reduce the data to right-censored data using the Cox PH classical estimation method and (2) the reduced formulation (see Appendix A of supplementary material available at *Biostatistics* online for details) that simplifies the data DI to single interval-censored data analyzed using the method of Pan (2000). Note that these reduced likelihood methods do not account for the impact of the covariate on U.

Table 1 shows the results of these simulations. The StEM algorithm provides less biased results than both univariate methods (midpoint and reduced formulation). In addition, we compared the estimated standard deviation of the parameter (STD) with the standard error of the simulation estimates (STE) and found that they were close for our approach. We note the large bias for the midpoint approach. The coverage probability of the 95% CI of the StEM is close to 0.95 and is certainly better than for the reduced likelihood methods.

Table 2. *Motivating dataset*: *duration of response in first-line metastatic breast cancer*

Estimator	$\beta(STD)$	p-value
Recorded time	−0.34 (0.18)	0.061
Midpoint for U and V	−0.29 (0.18)	0.145
Reduced formulation	−0.23 (0.23)	0.340
StEM	−0.30 (0.25)	0.260
StEM accounting for covariate on U	−0.17 (0.28)	0.278

STD = square root of estimated variance. p-value = Wald test for $\beta \neq 0$ in the distribution of T. StEM assumes an effect of treatment on U and T.

5. Data analysis

The motivating example is taken from a clinical trial in first-line metastatic breast cancer (Jassem *and others*, 2001). The trial studied the superiority of Taxol in combination with doxorubicin (AT) to the combination of 5-fluorouracil, doxorubicin, and cyclophosphamide (FAC) with respect to progression-free survival. Tumor measurements for the assessment of response and tumor progression were scheduled, per protocol, every 6 weeks during the treatment. The study randomized 267 subjects (134 subjects to the AT arm and 133 subjects to the FAC arm). Our analysis is based on 159 responders, with 87 responders observed in the AT arm and 72 in the FAC arm.

We applied the StEM algorithm to the data (with $m = 100$ as before), both ignoring and accounting for possible effect of the covariate on U. For comparison, we included the alternative approaches used in the simulation study, as well as the analysis using the recorded time (i.e. ignoring the interval-censored aspect) using classical Cox PH partial likelihood.

The results, given in Table 2, indicate that alternative approaches may lead to quite different conclusions, especially because of an underestimation of the STD. The results analyzed using the recorded time approach showed a borderline significant ($p = 0.061$) and favorable impact of treatment for AT. We see that the StEM approach, which accounts for all the variability of the measurements and for the covariate effect on U, shows a smaller, not statistically significant, effect favorable to AT.

6. Discussion

As outlined above, current methods for DI survival times are not adapted to the specificity of our dataset. Firstly, the methods for DI survival times actually allow only DR times when covariates are involved. Secondly, some methods require that the time intervals can be discretized, which was not suitable either for our dataset. Finally, the analysis shows the importance of accounting for the impact of the covariate on U in the estimation of the effect of treatment on duration of response. Therefore, the proposed StEM algorithm appears to be an appropriate method for analyzing duration of response type of data or any DI data in the context of clinical trials.

Finally, we wish to mention that the majority of the approaches that deal with DI survival times suppose that U and T are independent. In clinical trials, this lack of dependence may sometimes be questionable. This is a topic we wish to address in a subsequent paper. Programs have been written by the first author in R and are available upon request.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors would like to thank Professor Lupe Gomez for the interesting discussion and useful suggestions to improve the quality of this paper. We would also like to thank the Editor, Associate Editor, and the anonymous Referees for their valuable comments. *Conflict of Interest*: None declared.

APPENDIX A: CONSTRUCTION OF THE EXPECTED LOG LIKELIHOOD IN THE EM ALGORITHM

In this appendix, we detail the construction of the expectation of the log likelihood with respect to missing values, given the observed data and parameter values at the previous iteration, denoted by $Q_{k+1}(\theta|\theta^k)$.

At the $k+1$th iteration:

$$Q_{k+1}(\theta|\theta^k) = \int_{\substack{u_i,\forall i\in N\\ y_j,\forall j\in D}} \log L(\theta|y_i,\delta_i\forall i\in N)\,\mathrm{d}F_{U,T}(u_i,\forall i\in N, y_j,\forall j\in D|X_i,\mathcal{D},\theta^k),$$

where θ^k is the set of parameters estimated from the previous maximization step (M-step) of the EM algorithm. We can write, by the assumed independence of U and T and by the independence of the observations, that the joint distribution of U and T, given observed data, is given by

$$\begin{aligned}
&F_{U,T}(u_i,\forall i\in N, y_j\forall j\in D|X_i,\mathcal{D},\theta^k)\\
&\quad=\prod_{j\in D} F_{U,T}(u_j,y_j|X_j,u_j\in[u_{lj},u_{rj}],v_j\in[v_{lj},v_{rj}],\theta^k)\\
&\qquad\times\prod_{s\in C} F_U(u_s|X_s,u_s\in[u_{ls},u_{rs}],v_s\in[v_{ls},v_{rs}],\theta^k)\\
&\quad=\prod_{i=1}^{n} F_T(y_i|X_i,u_i,y_i\in[v_{li}-u_i,v_{ri}-u_i],\theta^k)^{\delta_i} F_U(u_i|X_i,u_i\in[u_{li},u_{ri}],v_i\in[v_{li},v_{ri}],\theta^k). \quad \text{(A.1)}
\end{aligned}$$

In (A.1), $F_T(y_i|X_i,u_i,y_i\in[v_{li}-u_i,v_{ri}-u_i],\theta^k)$ depends on the observed intervals of V and also on U despite the assumed independence between U and T, since the interval for y_i is constructed using the unobserved data u_i. For the formal derivation of $F_U(u_i|X_i,u_i\in[u_{li},u_{ri}],v_i\in[v_{li},v_{ri}],\theta^k)$, we first write the joint distribution of U and $U+T$ as

$$\Pr(U\in[u_{li},u_{ri}],U+T\in[v_{li},v_{ri}]|X,\theta^k)=\int_{u_{li}}^{u_{ri}}\int_{v_{li}}^{v_{ri}} f_U(u|X)f_T(v-u|X,\theta^k)\,\mathrm{d}v\,\mathrm{d}u,$$

by the assumed independence of U and T. Using the above equation, the conditional distribution of U given the observed data is written as

$$\begin{aligned}
&F_U(u_i|X_i,u_i\in[u_{li},u_{ri}],v_i\in[v_{li},v_{ri}],\theta^k)\\
&\quad=\int_{u_{li}}^{u_i} f_U(u|X_i,v_i\in[v_{li},v_{ri}],\theta^k)\,\mathrm{d}u\\
&\quad=\frac{1}{\mathrm{cst}}\int_{u_{li}}^{u_i} f_U(u|X_i)(F_T(v_{ri}-u|X_i,\theta^k)-F_T(v_{li}-u|X_i,\theta^k))\,\mathrm{d}u, \quad \text{(A.2)}
\end{aligned}$$

where $\mathrm{cst}=\int_{u_{li}}^{u_{ri}} f_U(u|X_i)(F_T(v_{ri}-u|X_i,\theta^k)-F_T(v_{li}-u|X_i,\theta^k))du$. Note that these derivations assume that none of the observed intervals of U and V overlap. We discuss the specific derivations when

overlapping is present in Appendix B. In summary, (A.2) justifies the notation that the conditional distribution of U given the observed data depends on the observed intervals for V, but also on the distribution of T (and the parameters θ^k).

Appendix B: Estimator of $F_U(u|\mathcal{D})$

Let us first assume that none of the observed intervals of U and V overlap. To obtain an estimator of $F_U(u|\mathcal{D})$, the estimators $\hat{f}_U(u|X_i)$ and $\hat{S}^k_{0,T}$ need to be plugged in (A.2).

We now describe how to obtain $\hat{f}_U(u|X_i)$. Note first that $\hat{f}_U(u|X_i)$ is fixed for all EM iterations since it depends on the observed $[u_{li}, u_{ri}]$, $i = 1, \ldots, n$. If f_U is independent of the covariate, its estimator is based on the Turnbull estimator. The parameters ψ represent the mass of the Turnbull estimator assigned to the constructed intervals, but the exact repartition of this mass within the intervals is not defined. Therefore, we assumed that the mass of $\hat{F}_U$ is uniformly distributed in these intervals because it leads to an invertible estimator of $F_U(u|\mathcal{D})$. If f_U depends on a covariate, a PH model is assumed and ψ represent the regression parameters and the baseline hazard parameters which are now estimated by the method of Pan (2000). Again, to obtain an invertible estimator of $F_U(u|\mathcal{D})$, we assume that the estimator is linear between simulated event times. Details of these procedures are given in Appendix B of supplementary material available at *Biostatistics* online. The second element of the integrand in (A.2) is $\hat{S}^k_{0,T}(t)$, which is obtained in the M-step. From the Breslow estimator of the baseline hazard in a Cox PH model, we know that $\hat{S}^k_{0,T}(t)$ is piecewise constant with jumps at event times. Hence, also $[\hat{S}^k_{0,T}(t)]^{\exp(\hat{\beta}^k X_i)}$ is piecewise constant. It follows from the piecewise linear form of $\hat{f}_U$ and $[\hat{S}^k_{0,T}(t)]^{\exp(\hat{\beta}^k X_i)}$ that the integral in (A.2) has a piecewise quadratic form, monotone increasing on its domain which can therefore be easily inverted to generate $\bar{u}_1^{k+1}, \ldots, \bar{u}_n^{k+1}$ $q = 1, \ldots, m$.

In case of overlapping intervals, we use the alternative to (A.2) that ensures that the integrand is evaluated on positive times:

$$\frac{\int_{u_{li}}^{u_i} \hat{f}_U(u|X_i)([\hat{S}^k_{0,T}(\max(v_{li}, u) - u)]^{\exp(\hat{\beta}^k X_i)} - [\hat{S}^k_{0,T}(v_{ri} - u)]^{\exp(\hat{\beta}^k X_i)})\, \mathrm{d}u}{\int_{u_{li}}^{u_{ri}} \hat{f}_U(u|X_i)([\hat{S}^k_{0,T}(\max(v_{li}, u) - u)]^{\exp(\hat{\boldsymbol{\beta}}^k X_i)} - [\hat{S}^k_{0,T}(v_{ri} - u)]^{\exp(\hat{\beta}^k X_i)})du}.$$

The calculation of the estimator of the transformed equation follows from the calculations above.

Appendix C: Construction of the variance estimator for β

The variance of the parameters is obtained from I_O, the information matrix of the observed likelihood (3.1). For simplicity, we will assume below that the covariate is unidimensional. Extension to a multidimensional covariate is straight forward. By the missing information principle given in Louis (1982), we have

$$\begin{aligned} I_O &= -\frac{\partial^2}{\partial\theta^2} \log L(\theta|u_i \in [u_{li}, u_{ri}], v_i \in [v_{li}, v_{ri}], \forall i \in N) \\ &= -\frac{\partial^2}{\partial\theta^2} \log L(\theta|y_i, \delta_i, \forall i \in N) \\ &\quad + \frac{\partial^2}{\partial\theta^2} \log f_{U,T}(u_i, \forall i \in N, y_j, \forall j \in D|\theta, u_i \in [u_{li}, u_{ri}], v_i \in [v_{li}, v_{ri}], \forall i \in N). \end{aligned} \tag{C.1}$$

Taking the expectation of (C.1) with respect to $u_i, i \in N,\ y_j, j \in D$ leads to

$$I_O = \underbrace{-\mathrm{E}_{\substack{u_i,\forall i\in N\\ y_j,\forall j\in D}}\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta|y_i,\delta_i,\ i=1,\ldots,n)\right]}_{A}$$

$$+\underbrace{\mathrm{E}_{\substack{u_i,\forall i\in N\\ y_j,\forall j\in D}}\left[\frac{\partial^2}{\partial\theta^2}\log f_{U,T}(u_i,\forall i\in N,y_j,\forall j\in D|\theta,u_i\in[u_{li},u_{ri}],v_i\in[v_{li},v_{ri}],\forall i\in N)\right]}_{B}.$$

We note that part A is the expected value of $I_T(\theta)$ (the information matrix of the right-censored data likelihood) with respect to the missing data. This matrix can be obtained by taking the second derivative of the right-censored data likelihood of a Cox PH model (see Klein and Moeschberger, 2003 or Goggins *and others*, 1999). The entries of this matrix are

$$\frac{\partial^2}{\partial\beta^2}\log L=\sum_{l=1}^{d}\left(-\lambda_l\sum_{j\in R(\tau_l)}X_j^2\exp(\beta X_j)\right),\quad \frac{\partial^2}{\partial\lambda_l^2}\log L=-1/\lambda_l^2,\quad \text{and}$$

$$\frac{\partial^2}{\partial\beta\partial\lambda_l}\log L=-\sum_{j\in R(\tau_l)}X_j\exp(\beta X_j),$$

where d is the number of events, $\tau_1,\ldots,\tau_d$ are the ordered event times, $R(\tau_l)$ is the risk set at event time τ_l, and λ_l is the hazard at time lth ordered event time.

Part B can be expressed (see Louis, 1982) as

$$B=-\mathrm{var}\left(\frac{\partial}{\partial\theta}\log L(\theta|y_i,\delta_i,\ i=1,\ldots,n)\right).$$

Note that $S(\theta)=(\partial/\partial\theta)\log L(\theta|y_i,\delta_i,\ i=1,\ldots,n)$ is the score vector of the right-censored likelihood, which can also be obtained easily (see Goggins *and others*, 1999). The entries of the score vector are

$$\frac{\partial}{\partial\beta}\log L=\sum_{l=1}^{d}\left[-\lambda_l\sum_{j\in R(\tau_l)}X_j\exp(\beta X_j)\right]+X_l,\quad \frac{\partial}{\partial\lambda_l}\log L=1/\lambda_l-\sum_{j\in R(\tau_l)}\exp(\beta X_j).$$

Goggins *and others* (1999) and Nielsen (2000) have proposed to estimate parts A and B at StEM iteration k by $\hat{A}=1/m\sum_{q=1}^{m}I_T(\hat{\theta})_q$ and $\hat{B}=1/m\sum_{q=1}^{m}S(\hat{\theta})_qS(\hat{\theta})_q'-1/m^2\sum_{q=1}^{m}S(\hat{\theta})_q(\sum_{q=1}^{m}S(\hat{\theta})_q)'$, where $I_T(\hat{\theta})_q$ is the information matrix and $S(\hat{\theta})_q$ is the score vector, both evaluated on dataset $\bar{y}_{1q}^k,\ldots,\bar{y}_{nq}^k$, with $\hat{\theta}=(\hat{\beta}^k,\hat{\lambda}_1^k,\ldots,\hat{\lambda}_d^k)$.

Further, Goggins *and others* (1999) propose to add $1/(m-1)B$ to account for the finite sampling in the StEM algorithm, which becomes negligible when m is sufficiently large.

An estimation for the information matrix of the parameters, based on observed data is therefore given by

$$\hat{I}_O=\frac{1}{m}\sum_{q=1}^{m}I_T(\hat{\theta})_q-\left(1+\frac{1}{m-1}\right)\left[\frac{1}{m}\sum_{q=1}^{m}S(\hat{\theta})_qS(\hat{\theta})_q'-\frac{1}{m^2}\sum_{q=1}^{m}S(\hat{\theta})_q\left(\sum_{q=1}^{m}S(\hat{\theta})_q\right)'\right],\qquad \text{(C.2)}$$

and the variance of $\hat{\beta}^k$ is obtained by inverting $\hat{I}_O$.

The estimator (C.2) assumes that the λ_l parameters are common to each of the m generated datasets, that is, that the event times are common. However, for each dataset, the event/censoring times are generated and are not identical. Given that we are only interested in the variance of the regression parameter $\hat{\beta}$, we can give an heuristic argument to justify why the estimator (C.2) is valid. Indeed, the variance of $\hat{\beta}$ is impacted by the component of the $\hat{I}_O$. These components depend only on the risk set and not directly on the time at which the hazards parameters are measured. Therefore, if our purpose is only to estimate the variance of $\hat{\beta}$, we can assume that the hazards parameters pertain to times that are common across datasets, and compute $\hat{I}_O$ from (C.2).

In the estimator (C.2), all $I_T(\hat{\theta})_q$ and $S(\hat{\theta})_q$ must have the same dimension. However, due to the piecewise constant nature of $\hat{S}_{0,T}^{k-1}$, $\bar{y}_j^k$, $j \in D$ are sampled from a finite set of event times. This leads to tied $\bar{y}_j^k$ and a set of parameter for $\hat{S}_{0,T}$ (which has one parameters for each event time) that can vary in size across m sampling. To obtain untied times and a parameter set that does not change in size (i.e. a constant d) across the generated dataset, we run an extra iteration $k+1$ in which $\bar{y}_j^{k+1}$ is sampled from $\hat{S}_{0,T}^k$, considered as piecewise linear.

References

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* **2**, 73–82.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45**, 1–11.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.

Goggins, W. B., Finkelstein, D. M. and Zaslavsky, A. M. (1999). Applying the Cox proportional hazards model for analysis of latency data with interval censoring. *Statistics in Medicine* **18**, 2737–2747.

Jara, A., Lesaffre, E., De Iorio, M. and Quintana, F. A. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Annals of Applied Statistics* **4**, 2126–2149.

Jassem, J., Pieńkowski, T., Płużańska, A., Jelic, S., Gorbunova, V., Mrsic-Krmpotic, Z., Berzins, J., Nagykalnai, T., Wigler, N., Renard, J. *and others*, for the Central & Eastern Europe. (2001). Doxorubicin and paclitaxel versus fluorouracil, doxorubicin, and cyclophosphamide as first-line therapy for women with metastatic breast cancer: final results of a randomized phase iii multicenter trial. *Journal of Clinical Oncology* **19**, 1707–1715.

Kim, M. Y., Gruttola, V. G. De and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* **49**, 13–22.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. New York: Springer.

Komárek, A. and Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association* **103**, 523–533.

Law, C. G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine* **11**, 1569–1578.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **44**, 226–233.

Nielsen, S. F. (2000). The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli* **6**, 457–489.

OLLER, R., GOMEZ, G. AND CALLE, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **32**, 315–326.

PAN, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**, 199–203.

PAN, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* **57**, 1245–1250.

SUN, L., KIM, Y.-J. AND SUN, J. (2004). Regression analysis of doubly censored failure time data using the additive hazards model. *Biometrics* **60**, 637–643.

SUN, J., LIAO, Q. AND PAGANO, M. (1999). Regression analysis of doubly censored failure time data with applications to AIDS studies. *Biometrics* **55**, 909–914.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)* **38**, 290–295.

[*Received January 11, 2013; revised April 19, 2013; accepted for publication April 20, 2013*]