BIOMETRICS 56, 1139–1144
December 2000

Semiparametric Regression Analysis of Interval-Censored Data

Els Goetghebeur

Ghent University, TWI, Krijgslaan 281-S9, B-9000 Ghent, Belgium,
email: els.goetghebeur@rug.ac.be

and

Louise Ryan

Harvard School of Public Health and Dana-Farber Cancer Institute,
44 Binney Street, Boston, Massachusetts 02115, U.S.A.

SUMMARY. We propose a semiparametric approach to the proportional hazards regression analysis of interval-censored data. An EM algorithm based on an approximate likelihood leads to an M-step that involves maximizing a standard Cox partial likelihood to estimate regression coefficients and then using the Breslow estimator for the unknown baseline hazards. The E-step takes a particularly simple form because all incomplete data appear as linear terms in the complete-data log likelihood. The algorithm of Turnbull (1976, *Journal of the Royal Statistical Society, Series B* **38,** 290–295) is used to determine times at which the hazard can take positive mass. We found multiple imputation to yield an easily computed variance estimate that appears to be more reliable than asymptotic methods with small to moderately sized data sets. In the right-censored survival setting, the approach reduces to the standard Cox proportional hazards analysis, while the algorithm reduces to the one suggested by Clayton and Cuzick (1985, *Applied Statistics* **34,** 148–156). The method is illustrated on data from the breast cancer cosmetics trial, previously analyzed by Finkelstein (1986, *Biometrics* **42,** 845–854) and several subsequent authors.

KEY WORDS: Breslow estimator; EM algorithm; Proportional hazards.

1. Introduction

Interval censoring arises in a variety of practical settings, e.g., in medical studies where subjects are periodically assessed for disease recurrence or progression. While straightforward parametric methods can be readily applied (see Odell, Anderson, and D'Agostino, 1992; Lindsey and Ryan, 1998), development of semiparametric methods for interval-censored data remains a topic of active research interest. Many authors have formulated the problem in terms of discrete survival, using methods for multinomially distributed data (cf., Turnbull, 1976; Finkelstein, 1986). A drawback is that these approaches do not reduce to standard survival analysis in settings where data are not interval censored. Furthermore, methods based on discrete survival can be numerically unstable, especially when covariates are involved, because discrete hazards are probabilities and hence need to be constrained between zero and one. To address this problem, Satten (1996) proposed a rank-based method, using Monte Carlo simulations, in combination with EM, which involves fitting standard Cox models for the complete-data setting. A disadvantage of the method, however, is its high computational demand. To address this, Satten, Datta, and Williamson (1998) revert to a parametric model for the imputation of missing exact or right-censored failure times. Goggins et al. (1998) also developed a less computationally demanding alternative to the Satten (1996) approach, though it still requires Monte Carlo simulation in the E-step of an EM algorithm. Pan (2000) proposes an approach based on multiple imputation of missing failure times based on Breslow's estimate of the survivorship function. A disadvantage of all these methods is the relatively *ad hoc* nature of the procedure used to impute missing data. The Pan (2000) method also suffers from a lack of guidelines on the choice of time points at which the survivorship function can jump. To avoid some of these problems, several authors have explored methods based on nonparametric smoothing of baseline hazards. Kooperberg and Clarkson (1997) use regression splines, while Betensky et al. (1999) use local likelihood smoothing. While these smoothing approaches lose the advantage of having Cox regression as a special case, they avoid some of the computational difficulties associated with the semiparametric techniques described above. In this paper, we propose an approach that retains some of the appealing features of the smoothing methods yet reduces to a standard Cox proportional hazards model in the absence of interval censoring. Our approach is also relatively tractable and has an approximate likelihood basis. Like Pan's approach, our method involves an appealing EM algorithm. The M-step estimates regression coefficients using Cox regression and baseline hazards using the Breslow estimator. The E-step requires estimating the risk

set sizes and number of events that occurred at each of a set of possible event times, identified using the algorithm of Turnbull (1976). Our E-step is computationally simple and needs no iterations. When the data are right censored, the method reduces exactly to the standard Cox proportional hazards analysis and the algorithm corresponds to the one proposed by Clayton and Cuzick (1985). Several approaches are discussed for estimating the variance of estimated parameters. Of these, the multiple imputation variance of Rubin and Schenker (1991) appears to have the best small-sample properties.

In Section 2, we introduce notation and the proportional hazards model assumed to underlie the interval-censored data. Section 3 motivates the approximate likelihood used as the basis for our approach and discusses estimation for the complete-data setting. Section 4 describes the steps of the EM algorithm and discusses the choice between candidate variance estimators. A small simulation study illustrates the properties of our approach. Section 5 applies the method to the breast cancer cosmesis data presented by Finkelstein (1986), and we end with a discussion.

2. Notation

For $i = 1, \ldots, n$, let T_i represent time until the event of interest for individual i and suppose T_i has intensity process defined by

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T_i < t + \Delta t \mid T_i \geq t) Y_i(t) = \lambda_i(t) Y_i(t),$$

where $Y_i(t)$ indicates whether subject i remains at risk at time t. The dependence of the intensity functions on i allows for the inclusion of individual covariates, such as exposure level. Let $\mathbf{Z}_i$ represent a vector of covariates for the ith individual. Then a standard proportional hazards assumption implies

$$\lambda_i(t) = \lambda(t) \exp(\boldsymbol{\tau}' \mathbf{Z}_i), \tag{1}$$

where $\boldsymbol{\tau}$ represents a vector of unknown regression coefficients.

We consider the setting where some or all of the data are interval censored, i.e., the event time T_i for person i is not directly observable but instead is known only to have occurred within an interval $[L_i, U_i]$. The censoring mechanism is assumed to be one that yields ignorably coarsened data (Heitjan and Rubin, 1991).

The primary problem in fitting model (1) to interval-censored data is that the standard partial likelihood formulation is not easily adapted (Gill, 1992). In the next section, we present an approximate complete-data likelihood that yields the standard Cox proportional hazards analysis in the usual right-censored survival case. The approximate likelihood facilitates an EM approach to the proportional hazards regression analysis of interval-censored survival data.

3. An Approximate Likelihood for Right-Censored Data

Assume for now that T_i is either observed exactly or noninformatively right censored. Let δ_i be a censoring indicator taking the value one if the event time T_i is observed exactly and zero if T_i is right censored. Then the full-data likelihood is proportional to

$$L = \prod_i \{\lambda(t_i) \exp(\boldsymbol{\tau}' \mathbf{Z}_i)\}^{\delta_i} \times \exp\left\{ - \int_0^{t_i} \lambda(u) \exp(\boldsymbol{\tau}' \mathbf{Z}_i) Y_i(u) du \right\}, \tag{2}$$

where $Y_i(t)$ indicates whether individual i is still at risk for an event at time t.

Instead of replacing (2) by a partial likelihood, as would be done in the usual setting, suppose we assume that the T_i's follow a piecewise exponential model with hazard rates λ_k, which are constant on a fine grid of unit-width intervals. We assume that the time unit is chosen finely enough that tied failure times are unlikely to occur. Our approximation arises because we assume that each observed failure or censoring time is rounded up to the endpoint of the interval within which it occurred. Let Y_{ik} indicate whether individual i is at risk in interval k and let λ_k represent the baseline hazard attributed to the kth interval. Let $N_i(t)$ be a counting process indicating whether or not the event has occurred by time t in individual i and, with a slight abuse of notation, let dN_{ik} indicate whether or not the event occurred in the kth interval for that subject. Under these assumptions, L can be rewritten as

$$\tilde{L} = \prod_i \prod_k \left[\{\lambda_k \exp(\boldsymbol{\tau}' \mathbf{Z}_i)\}^{dN_{ik}} \exp\left\{ -\lambda_k \exp(\boldsymbol{\tau}' \mathbf{Z}_i) Y_{ik} \right\} \right], \tag{3}$$

where the products are over intervals $k = 1, \ldots, K$ and individuals $i = 1, \ldots, N$. The corresponding log likelihood is

$$\begin{aligned} l &= \log(\tilde{L}) \\ &= \sum_i \sum_k \left\{ \log(\lambda_k) dN_{ik} + \boldsymbol{\tau}' \mathbf{Z}_i dN_{ik} - \lambda_k \exp(\boldsymbol{\tau}' \mathbf{Z}_i) Y_{ik} \right\}. \end{aligned} \tag{4}$$

Score equations derived from (4) for λ_k and $\boldsymbol{\tau}$ are

$$\frac{\partial l}{\partial \lambda_k} = \sum_i \left\{ \frac{1}{\lambda_k} dN_{ik} - \exp(\boldsymbol{\tau}' \mathbf{Z}_i) Y_{ik} \right\} = 0 \tag{5}$$

and

$$\frac{\partial l}{\partial \boldsymbol{\tau}} = \sum_k \sum_i \left\{ \mathbf{Z}_i dN_{ik} - \lambda_k \mathbf{Z}_i \exp(\boldsymbol{\tau}' \mathbf{Z}_i) Y_{ik} \right\} = 0. \tag{6}$$

Note that, for fixed and known $\boldsymbol{\tau}$, the solution to equation (5) is

$$\hat{\lambda}_k(\boldsymbol{\tau}) = \frac{\sum_i dN_{ik}}{\sum_i \exp(\boldsymbol{\tau}' \mathbf{Z}_i) Y_{ik}}. \tag{7}$$

Because $\hat{\lambda}_k(\boldsymbol{\tau})$ is nonzero only for intervals in which an event occurs, this turns out to be equivalent to the Breslow estimator of the baseline hazard (see Andersen et al., 1993; also Breslow's discussion of Cox, 1972), which in the complete data case yields the Nelson–Aalen estimator for the survivor function. A profile log likelihood can be found by substituting $\hat{\lambda}_k(\boldsymbol{\tau})$ into the expression at (4). Differentiating with respect to $\boldsymbol{\tau}$ yields the following profile likelihood score equation, which is equivalent to simply substituting $\hat{\lambda}_k(\boldsymbol{\tau})$ from (7) into (6),

$$\sum_k \left\{ \sum_i (\mathbf{Z}_i - \bar{\mathbf{Z}}_k) dN_{ik} \right\} = 0, \tag{8}$$

where $\bar{\mathbf{Z}}_k$ is the weighted average of the $\mathbf{Z}$'s among the individuals still at risk for the event at time t_k,

$$\bar{\mathbf{Z}}_k = \frac{\sum_i \mathbf{Z}_i \exp(\boldsymbol{\tau}'\mathbf{Z}_i) Y_{ik}}{\sum_i \exp(\boldsymbol{\tau}'\mathbf{Z}_i) Y_{ik}}.$$

The negative derivative of the profile score (8) can be inverted to yield an estimate of the variance of $\hat{\boldsymbol{\tau}}$. An equivalent result can be obtained as the $(\boldsymbol{\tau}, \boldsymbol{\tau})$ element of the inverse information matrix based on differentiating (5) and (6), which is in turn equivalent to the usual variance estimator based on the standard partial likelihood. This same result was observed by Clayton and Cuzick (1985) and is consistent with many examples in the literature where estimation of an infinite dimensional nuisance parameter does not disturb the usual likelihood-based properties of estimates of a finite dimensional parameter of interest (see Murphy (1994) for an example based on proportional hazards frailty models and Murphy and Van der Vaart (1999) for a more general development).

In addition to its connection with the standard Cox partial likelihood, an advantage of working with the approximation (3) is the linearity of the corresponding score equations (5) and (6) in the counting process terms N and Y. As we will see in the next section, this linearity facilitates application of an EM algorithm for the interval-censored setting.

4. Interval-Censored Data

Following the notation introduced in Section 2, suppose now that, instead of being directly observed, the event time for individual i is known only to lie in an interval $[L_i, U_i]$. Some observations may be observed exactly ($L_i = U_i$) while others may be right censored ($U_i = \infty$). To facilitate application of an EM algorithm, we consider exact or right censoring to be the complete-data setting, so the approximate log likelihood (4) can be thought of as the complete-data log likelihood.

4.1 *E-Step*

At the E-step, we need to find the expected value of the complete-data (log-likelihood) score conditional on the observed data. Because expression (4) is linear in the components of the counting process defined earlier as well as in the process Y_{ik}, the E-step simply involves calculating the conditional expectation of Y_{ik} and dN_{ik} given the observed data, i.e., we need the conditional expectation of the number at risk for events at time t_k and the expected number of events at t_k. Let these conditional expectations be r_{ik} and e_{ik}, respectively. Then the expected complete-data score equations for λ_k and $\boldsymbol{\tau}$ are, respectively,

$$\sum_i \left\{ \frac{1}{\lambda_k} e_{ik} - \exp(\boldsymbol{\tau}'\mathbf{Z}_i) r_{ik} \right\} = 0$$

and

$$\sum_k \sum_i \left\{ \mathbf{Z}_i e_{ik} - \lambda_k \mathbf{Z}_i \exp(\boldsymbol{\tau}'\mathbf{Z}_i) r_{ik} \right\} = 0.$$

For non–right-censored observations ($U_i \neq \infty$), the conditional expectation of dN_{ik} is given by

$$e_{ik} = \begin{cases} \dfrac{p_{ik}}{\sum_{l_i \le t_l \le u_i} p_{il}} & \text{if } t_k \in [l_i, u_i], \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where

$$p_{ik} = \lambda_k \exp(\boldsymbol{\tau}'\mathbf{Z}_i) \left[\exp\left\{ -\sum_{j=1}^{k} \lambda_j \exp(\boldsymbol{\tau}'\mathbf{Z}_i) \right\} \right].$$

For right-censored observations ($U_i = \infty$), we set $e_{ik} = 0$ for all values of k.

The conditional expectation of Y_{ik} is given by

$$r_{ik} = \sum_{l \ge k} e_{il}\delta_i + I(t_k \le l_i)(1 - \delta_i), \tag{10}$$

so for right-censored observations, $r_{ik} = 1$ for $t_k \le l_i$.

4.2 *M-Step*

The solution to the expected score equation for $\lambda_k(\boldsymbol{\tau})$ takes the form

$$\hat{\lambda}_k(\boldsymbol{\tau}) = \frac{\sum_i e_{ik}}{\sum_i \exp(\boldsymbol{\tau}'\mathbf{Z}_i) r_{ik}}, \tag{11}$$

while $\hat{\boldsymbol{\tau}}$ is the solution to

$$\sum_k \sum_i \{\mathbf{Z}_i - \bar{\mathbf{Z}}_k(\boldsymbol{\tau})\} e_{ik} = 0, \tag{12}$$

where $\bar{\mathbf{Z}}_k(\boldsymbol{\tau})$ is the weighted average

$$\bar{\mathbf{Z}}_k(\boldsymbol{\tau}) = \frac{\sum_i \mathbf{Z}_i \exp(\boldsymbol{\tau}'\mathbf{Z}_i) r_{ik}}{\sum_i \exp(\boldsymbol{\tau}'\mathbf{Z}_i) r_{ik}}.$$

After specifying initial values, the algorithm proceeds by iterating between the calculation of e_{ik} and r_{ik} from (9) and (10) and then updating the estimates of λ_k and $\boldsymbol{\tau}$ using (11) and (12). Iteration should proceed until all parameter estimates have converged. A good choice for starting values may be obtained by assuming that event times occur at the midpoints of censoring intervals and then applying a standard Cox proportional hazards model.

4.3 *Choice of Time Points*

Although the algorithm will work if all the λ_k's are viewed as unknown parameters to be estimated, the algorithm will converge to a single set of hazard masses and more quickly if one identifies in advance the times k for which λ_k will be estimated with a nonzero value. Close examination of the approximate likelihood (3) reveals that the same logic used by Turnbull (1976) can be applied to identify equivalence classes within which the observed data likelihood will be insensitive to any redistribution of the probability mass within that equivalence class. A convenient way to identify a set of appropriate t_k's is then as follows. First, sort the set of all observed L_i and U_i values. Based on this ordering, whenever an observed U_j is immediately preceded by an L_i, the corresponding interval $[L_i, U_j]$ is an equivalence class containing positive mass. We have chosen the midpoint of this interval as the location for the mass per equivalence class in our application.

4.4 *Variance Estimation*

Several factors complicate the variance estimation for the parameters obtained following our proposed method. First, the dimension of $\boldsymbol{\lambda}$ is large and likely to exceed the sample size. This problem is not unique but is common to many semiparametric approaches to interval-censored data.

Table 1
Simulation results, groups of size 50

Empirical	Min	Q1	Median	Mean	Q3	Max	SD
Point estimate	−0.05161	0.5214	0.6748	0.6841	0.855	1.28	0.2435
Robust sandwich estimate SD	0.2329	0.2742	0.2939	0.2994	0.3152	0.4206	0.0352
Imputation based estimate SD	0.2295	0.2457	0.2507	0.2519	0.2577	0.2834	0.0093

Addressing the problem rigorously is difficult (Huang and Wellner, 1995); hence, we adopt the commonly used approach of assuming on heuristic grounds that the usual likelihood derivations for the parameter of interest will remain valid. The simulation results at the end of this section support this.

The fact that this method is based on an EM algorithm also complicates the calculation of the variance of our estimator. Several approaches to calculating the observed information matrix in an EM context have been proposed (Louis, 1982; Meng and Rubin, 1991). In our case, the large number of parameters makes the SEM approach of Meng and Rubin (1991) computationally intractable. The Louis approach is more appealing because the correction term to be subtracted from the conditional expectation of the complete-data information matrix involves only calculation of the conditional expectation of polynomials of the second degree in Y_{ik} and dN_{ik}. Still, this involves tedious algebra and computation to obtain the appropriate terms. Another alternative was suggested by Mieltjen (1992), who observed that the EM algorithm can be thought of as solving the expected complete-data score equation,

$$\sum_{i=1}^{n} \mathbf{U}_i = 0,$$

where $\mathbf{U}_i$ is the expected value of the ith individual's contribution to the complete data score, given their observed data. Hence, the information matrix, or the variance of the score, can be empirically estimated by

$$\boldsymbol{\Sigma} = \sum_{i=1}^{n} \mathbf{U}_i \mathbf{U}_i^{\mathrm{T}}.$$

In addition to its disadvantage in terms of efficiency, the Mieltjen approach will be problematic in our setting due to the dimension of the parameter space. A more accurate, albeit more computationally demanding, alternative is to further employ the theory of estimating equations, which would suggest estimating the variance of the score by $\mathbf{D}^{-1}$ where $\mathbf{D}$ is the matrix of derivatives of the expected complete-data score,

$$\mathbf{D} = \sum_{i=1}^{n} \frac{\partial \mathbf{U}_i}{\partial \boldsymbol{\theta}} = 0, \tag{13}$$

where $\boldsymbol{\theta}$ represents the vector of unknown parameters, $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_k, \tau)$. A robust alternative is $\mathbf{D}^{-1}\boldsymbol{\Sigma}(\mathbf{D}^{-1})^{\mathrm{T}}$. Further details can be found in Elashoff and Ryan (1996), who also discuss ways to simplify the calculations. Related ideas have been discussed by Heyde and Morton (1996). An appendix providing details on how to calculate this is available from the authors on request.

A simpler variance estimator, which also turns out to have good small-sample properties, is one based on multiple imputation. Following Rubin and Schenker (1991), we impute possible values for the interval-censored data, using the conditional sampling mechanism described in Section 4.1, after the final step of the EM algorithm is completed. Each imputed data set yields a naive proportional hazards point and variance estimate for τ. A variance estimate for the EM estimator can then be found as a weighted sum of the empirical variance of the imputation estimates and the mean of the imputation variances. The weights are $1 + 1/m$ and 1, respectively, where m is the number of imputations that is used. In our case, we chose $m = 20$.

4.5 *Simulated Example*

To illustrate the potential of our approach, we generated 200 samples as follows. For $n = 50$ independent subjects on each arm, exponential survival data and independent right-censoring times were generated with hazards 1/30 and 1/90, respectively, in the control arm. In the exposed arm, the survival hazard is multiplied by a factor two, i.e., $\tau = \log(2) = 0.6931$.

We generated up to eight between-visit times from independent normal distributions with mean four and variance one. If the original survival time was not right censored, we let it be interval censored by the neighboring constructed visits. If an observed survival time thus fell beyond the last visit time, it was in turn right censored at that last visit time. A staggered entry into the study is simulated as a normal with mean 0.43 and standard deviation 0.05 (left censored at zero). The results of applying our method to the simulated data are summarized in Table 1.

The results suggest that the method provides an unbiased estimate of the true exposure effect. Standard error estimates based on the asymptotic approximation appear to be too large. In contrast, the 95% confidence interval derived from the imputation based variance estimate covered the true $\tau = \log(2)$ in 95.5% of the simulations.

5. Example

We reanalyze the breast cancer data from Finkelstein (1986), where the goal was to compare time to cosmetic deterioration between 46 women who received radiotherapy alone and 48 women who received radiotherapy plus chemotherapy. As a single covariate, an indicator of radiotherapy plus chemotherapy was used in model (1). Of the 94 women in the study, 56 experienced cosmetic deterioration and 38 did not (and hence were right censored for the purposes of analysis). Because detection of deterioration required a clinic visit, the 56 event times were interval censored, usually within a period of about 6 months. Applying our proposed algorithm (and iterating until there was less than 0.0001 absolute change in

Table 2
Treatment effect for breast cosmesis data

Model	Estimate	Standard error
Finkelstein (1986)	0.791	0.288
Exponential	0.742	0.277
Satten (1996)	0.890	0.297
Satten et al. (1998)	0.878	0.294
Goggins et al. (1998)	1.450	0.371
Piecewise (eight intervals)	0.930	0.287
Local likelihood	1.053	0.270
Profile likelihood + EM	0.784	0.280

any parameter estimate) yields an estimated τ of 0.784, with a standard deviation of 0.280, yielding a two-sided p-value for τ of 0.019. In comparison, Finkelstein (1986) found an estimator for τ equal to 0.791 with a standard error of 0.2880. In Table 2, we compare our results with those found by other methods.

The proposed approach yields a result in line with the other approaches and with a fairly narrow confidence interval, which the simulations suggest is not deceivingly narrow.

Figure 1 shows the predicted survival curves within the two groups according to this model. For comparison, we also give the curves obtained when fitting our model separately in the two groups (without any covariates). The two sets of curves are quite close, lending support to the proportional hazards assumption underlying the analysis except perhaps at the very start of the trial.

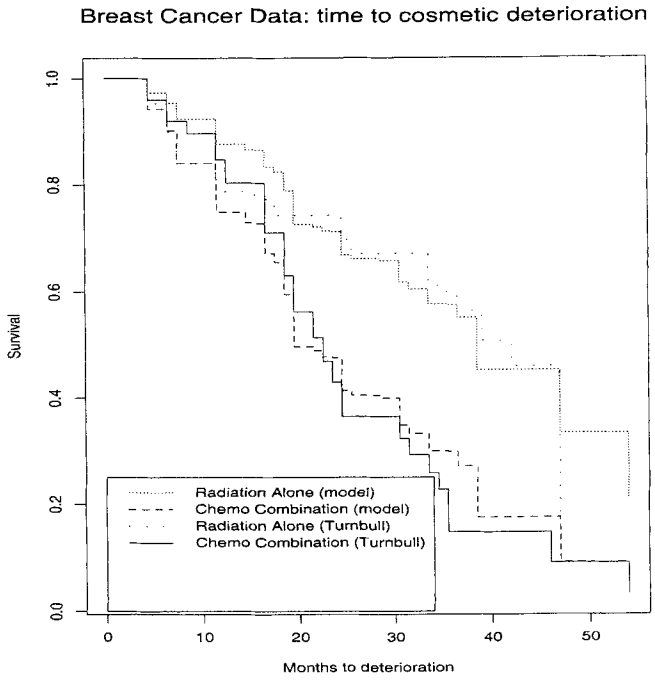

Figure 1. Estimated survival curves in separate groups and model based.

6. Discussion

We have proposed the use of an approximate likelihood for the analysis of continuous, interval-censored survival data. Parameters are estimated using an EM algorithm with appealingly simple steps. The M-step involves fitting a standard Cox proportional hazards model to estimate regression parameters, then using the Breslow estimator for baseline hazards. The E-step involves calculating the number of individuals at risk and the expected number of events at mass points identified using the algorithm proposed by Turnbull (1976). One can estimate the variance of estimated regression coefficients using an asymptotic approximation based on the theory of estimating equations. However, simulations suggest that multiple imputation will provide a better variance approximation, especially in small samples. In the special case when data are exact or right censored, our approach yields the same results as the standard proportional hazards model based on a partial likelihood.

Conceptually, to obtain a nonparametric maximum likelihood estimator (NPMLE) for the hazard, we approximate the nonparametric baseline hazard by piecewise constant hazards over a very fine grid of time points. The approximate model, like the original one, is overparameterized given only right-censored data. Interest lies, however, primarily in the finite dimensional parameter expressing treatment effect. Murphy and Van der Vaart (1999) have shown that, in general, the presence of infinite dimensional nuisance parameters need not preclude maximum profile likelihood estimators from inheriting usual properties, including a consistent variance estimator. Interval censoring adds a missing data problem, which is handled by first letting the interval-censored failure times coincide with a set of time points, one per Turnbull equivalence class, and then imbedding the profile likelihood approach in an EM algorithm.

Several features lend support to the hope that the usual properties could eventually be rigorously justified. First, we have shown how the analysis reduces to the standard Cox proportional hazards analysis in the special case of exact or right-censored data. Further, Huang (1999) showed that the NPMLE converges weakly to a Gaussian process for a mix of interval-censored data and exactly observed failure times. In a single group, our estimate reduces to the NPMLE, while in the case of two groups, Figure 1 shows that these estimators are close. Finally, Table 2 shows that the proposed estimator compares well with other estimators currently applied to this problem. However, as in those other cases, further work is needed to understand the theoretical properties of the proposed estimator in more detail.

Acknowledgements

This work was supported by grant CA48061 from the NCI and a travel grant CRG 950648 from NATO. We are grateful to Dr Satten for the results of his approach applied to the cosmesis data and to the associate editor for helpful comments.

Résumé

Nous proposons une approche semi-paramétrique pour l'analyse dumodèle de risques proportionnels appliqué à des données censurées par intervalle. Un algorithme EM basé sur une vraisemblance approximée conduit à une étape EM qui comporte la maximisation de la vraisemblance partielle de Cox, et l'utilisation de l'estimateur de Breslow pour les risques de base inconnus. L'étape E prend une forme particulièrement simple parce que les données incomplètes apparaissent comme des termes linéaires de la vraisemblance complète. L'algo-

rithme de Turnbull (1976) est utilisé pour déterminer les temps auxquels la fonction de risque peut prendre une masse positive. Des imputations multiples produisent une estimation de la variance facile à calculer et qui apparaît plus sûre que les méthodes asymptotiques pour des échantillons de taille modérée. Pour des données censurées par intervalle l'approche se réduit à l'analyse standard du modèle de Cox, tandis que l'algorithme est celui suggéré par Clayton et Cuzick (1985). La méthode est illustrée sur des données provenant d'un essai cosmétique sur le cancer du sein, déjà analysé par Finkeslstein (1986) et d'autres auteurs.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

Betensky, R. A., Lindsey, J. C. , Ryan, L. M., and Wand, M. P. (1999). Local EM estimation of the hazard function for interval-censored data. *Biometrics* **55,** 238–245.

Clayton, D. and Cuzick, J. (1985). The EM algorithm for Cox's regression model using GLIM. *Applied Statistics* **34,** 148–156.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Elashoff, M. and Ryan, L. (2001). An EM algorithm to adjust for missing data in estimating equations. *Journal of Computational and Graphical Statistics,* in press.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845–854.

Gill, R. D. (1992). Marginal partial likelihood. *Scandinavian Journal of Statistics* **19,** 133–137.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A., and Zaslavsky, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* **54,** 1498–1507.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics* **19,** 2244–2253.

Heyde, C. C. and Morton, R. (1996). Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society, Series B* **58,** 317–327.

Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval censored data. *Statistica Sinica* **9,** 501–519.

Huang, J. and Wellner, J. (1995). Asymptotic normality of the NPMLE of linear functionals for interval-censored data, case 1. *Statistica Neerlandica* **49,** 153–163.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics* **53,** 1485–1494.

Lindsey, J. and Ryan, L. (1998). Methods for interval censored data. Tutorial in biostatistics. *Statistics in Medicine* **17,** 219–138.

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

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86,** 899–909.

Mielijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Series B* **51,** 127–138.

Murphy, S. (1994). Asymptotic theory for the frailty model. *Annals of Statistics* **23,** 182–198.

Murphy, S. A. and Van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli* **5,** 381–412.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992). Maximum likelihood estimation for interval censored data using a Weibull-based accelerated failure time model. *Biometrics* **48,** 951–959.

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

Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-care data bases: An overview and some applications. *Statistics in Medicine* **10,** 585–598.

Satten, G. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83,** 355–370.

Satten, G., Datta, S., and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* **93,** 318–327.

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

Received March 1997. *Revised April* 2000.
Accepted May 2000.