

## A penalized likelihood approach for mixture cure models

Fabien Corbière<sup>1,2,\*</sup>, Daniel Commenges<sup>1,2</sup>, Jeremy M. G. Taylor<sup>3</sup> and Pierre Joly<sup>1,2</sup>

<sup>1</sup>*INSERM U897 Biostatistique, Bordeaux F-33076, France*

<sup>2</sup>*Université Victor Ségalen Bordeaux 2, Bordeaux F-33076, France*

<sup>3</sup>*Department of Biostatistics, 1420 Washington Heights, University of Michigan, Ann Arbor, MI 48109, U.S.A.*

### SUMMARY

Cure models have been developed to analyze failure time data with a cured fraction. For such data, standard survival models are usually not appropriate because they do not account for the possibility of cure. Mixture cure models assume that the studied population is a mixture of susceptible individuals, who may experience the event of interest, and non-susceptible individuals that will never experience it. Important issues in mixture cure models are estimation of the baseline survival function for susceptibles and estimation of the variance of the regression parameters. The aim of this paper is to propose a penalized likelihood approach, which allows for flexible modeling of the hazard function for susceptible individuals using M-splines. This approach also permits direct computation of the variance of parameters using the inverse of the Hessian matrix. Properties and limitations of the proposed method are discussed and an illustration from a cancer study is presented. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: survival analysis; mixture cure models; risk function; penalized likelihood; splines

### 1. INTRODUCTION

Models for survival analysis typically assume that everybody in the study population is susceptible to the event of interest and will eventually experience this event if the follow-up is sufficiently long. However, in some diseases there is strong biological evidence that, when considering endpoints other than natural death, the event would never occur for some fraction of the subjects. The assumption that some individuals will never experience the event may also be based solely on empirical considerations, such as the presence of a large number of long-term survivors. The presence of ‘non-susceptible’ individuals (or ‘immune’ or ‘cured’ regarding the event of interest) is suggested by a Kaplan–Meier plot of the marginal survival distribution function, which shows a long and stable plateau with heavy censoring at the right extreme. In recent years, there has been

\*Correspondence to: Fabien Corbière, Department of Large Animals Medicine, National Veterinary School, 23 Chemin des Capelles, 31076 Toulouse, France.

†E-mail: f.corbiere@envt.fr

an increasing interest in modeling survival data with long-term survivors. Failing to account for such cured subjects would lead to overestimation of the survival of susceptible subjects. Mixture cure models assume that the studied population is a mixture of susceptible (uncured) individuals, who may experience the event of interest, and non-susceptible (cured) individuals, who will never experience it. This approach allows us to estimate simultaneously whether the event of interest will occur, which is called incidence in the related literature, and when it will occur, given that it can occur, which is called latency.

An important issue arising from mixture cure models is estimation of the baseline survival function for uncured individuals. Fully parametric distributions were first proposed, including exponential [1–3], lognormal [4, 5], Weibull [6], Gompertz [7, 8] and Burr XII [9]. However, fully parametric mixture cure models may not be flexible enough when analyzing biological data, since they involve strong assumptions about the survival function of uncured patients. The generalized  $F$  distribution was proposed by Peng *et al.* [10], which makes less distributional assumptions, but computational difficulties may arise. Another approach is based on a non-parametric estimate of the survival function of uncured individuals, and this defines the area of semiparametric mixture cure models. However, unlike the standard Cox proportional hazards (PH) model [11], the baseline survival distribution function in semiparametric PH mixture cure models cannot be eliminated as a nuisance parameter in the likelihood function and an estimation method more complicated than maximizing the partial likelihood is needed. A semiparametric PH specification for the time to recurrence of the disease was proposed by Kuk and Chen [12], but a Monte Carlo approximation of the likelihood was required, which complicates parameter estimation. Assuming no covariates affect the latency, Taylor [13] used a weighted Kaplan–Meier estimator to estimate the failure time distribution and proposed an EM-algorithm. Peng and Dear [14] further extended this work by allowing the failure time distribution to depend on explanatory variables, but Monte Carlo techniques were still required to estimate the variance of the estimates. In the same way, Sy and Taylor [15, 16] use the EM-algorithm to estimate the parameters, but proposed to estimate the asymptotic variance of the estimator by the inverse of the observed information matrix, based on the observed marginal likelihood function using the discrete PH model. Methods based on the EM-algorithm adopted the zero-tail constraint proposed by Taylor [13] to avoid identifiability problems. This constraint states that the conditional survival function is set to zero for censored times greater than the largest uncensored event time, which ensures a proper estimate. However, it implies that individuals with survival times greater than the last event are all considered as non-susceptible or cured, which may appear to be a strong assumption, especially in the case of poor follow-up beyond the period when events occur or in the case of heavy censoring. Peng [17] further proposed different tail-completion methods to allow the conditional survival estimate to decrease slowly to zero after the largest event time. Based on simulation studies, he also showed that the zero-tail constraint could lead to an overestimated cured fraction.

Another drawback of semiparametric approaches is that a smooth estimate of the hazard function, which often has a meaningful interpretation, is not available. An alternative approach is to define the estimator of the conditional hazard function non-parametrically as the function, which maximizes a penalized likelihood. The solution is then approximated using splines [18]. Such an approach has been proposed in standard survival models and multi-state models under the PH assumption for right-censored and left-truncated data [19, 20]. In this paper we shall apply the penalized likelihood approach to mixture cure models. This method allows us to obtain flexible and smooth estimates of the conditional and marginal hazard and survival functions and does not require a tail constraint. The asymptotic variances of the estimators are obtained using the inverse of the

information matrix. The model is presented in the next section. The penalized likelihood approach and the estimation procedure is described in Section 3. In Section 4 a simulation study is shown. An illustrative example is provided in Section 5.

## 2. THE MODEL

### 2.1. Mixture cure models

Let  $U$  be the unobserved indicator denoting whether an individual is susceptible ( $U = 1$ ) or non-susceptible ( $U = 0$ ) to the event of interest and  $T$  is a non-negative random variable denoting the failure time of interest. Taking into account the possibility that a fraction of the population can be considered as immune or cured, the marginal, population-based survival and probability density functions are given by

$$S_{\text{pop}}(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})S(t|U = 1, \mathbf{x}) + 1 - \pi(\mathbf{z}) \quad (1)$$

$$f_{\text{pop}}(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})f(t|U = 1, \mathbf{x}) \quad (2)$$

where  $S_{\text{pop}}(t|\mathbf{x}, \mathbf{z})$  and  $f_{\text{pop}}(t|\mathbf{x}, \mathbf{z})$  are the unconditional (marginal) survival and probability density functions of  $T$  for the entire population, respectively.  $S(t|U = 1, \mathbf{x}) = P(T > t|U = 1, \mathbf{x})$  is the survival function for susceptible individuals given a covariate vector  $\mathbf{x} = (x_1, \dots, x_p)'$ , and  $\pi(\mathbf{z}) = P(U = 1|\mathbf{z})$  is the probability of being susceptible given a covariate vector  $\mathbf{z} = (z_1, \dots, z_q)'$ , which may include the same covariates as  $\mathbf{x}$ . Mixture models assume that  $S(t|U = 1, \mathbf{x})$  is proper (i.e. total mass equals 1), so that  $\lim_{t \rightarrow \infty} S(t|U = 1, \mathbf{x}) = 0$  and  $\lim_{t \rightarrow \infty} S_{\text{pop}}(t|\mathbf{x}, \mathbf{z}) = 1 - \pi(\mathbf{z})$ . The zero-tail constraint proposed by Taylor [13] assumes that there is a  $t^*$  such that for all  $t \geq t^*$ ,  $S(t|U = 1, \mathbf{x}) = 0$ . Usually one takes  $t^* = t_\kappa$ , where  $t_\kappa$  is the last observed failure time. When  $\pi(\mathbf{z}) = 1$  for all  $\mathbf{z}$ , i.e. when there is no cured fraction, the mixture cure model reduces to the standard survival model.

The proportion of cured patients can be modeled by the usual binary regression models. Farewell [6], Kuk and Chen [12], Peng and Dear [14] and Sy and Taylor [15] all used a logistic regression model with logit link,  $\text{logit}(\pi(\mathbf{z})) = \boldsymbol{\beta}'\mathbf{z}$ , where  $\boldsymbol{\beta}$  is the vector of regression parameters associated with  $\mathbf{z}$ , including an intercept. Other regression models include the probit link and the complementary log–log link [21].

In parametric mixture cure models,  $S(t|U = 1)$  is specified by a few parameters, whereas in semiparametric mixture cure models it is left arbitrary. In PH models, the conditional distribution of  $T$  is modeled by

$$\begin{aligned} S(t|U = 1, \mathbf{x}) &= S_0(t|U = 1)^{\exp(\boldsymbol{\gamma}'\mathbf{x})} \\ &= \exp\left(-\exp(\boldsymbol{\gamma}'\mathbf{x}) \int_0^t \lambda_0(v|U = 1) dv\right) \end{aligned} \quad (3)$$

where  $S_0(t|U = 1)$  and  $\lambda_0(t|U = 1)$  are the baseline conditional survival and hazard functions, respectively. The marginal, population based, hazard function is then given by

$$\begin{aligned} \lambda_{\text{pop}}(t|x, z) &= \frac{f_{\text{pop}}(t|\mathbf{x}, \mathbf{z})}{S_{\text{pop}}(t|\mathbf{x}, \mathbf{z})} \\ &= \frac{\pi(\mathbf{z})\lambda_0(t|U = 1)\exp(\boldsymbol{\gamma}'\mathbf{x})S_0(t|U = 1)^{\exp(\boldsymbol{\gamma}'\mathbf{x})}}{\pi(\mathbf{z})S_0(t|U = 1)^{\exp(\boldsymbol{\gamma}'\mathbf{x})} + 1 - \pi(\mathbf{z})} \end{aligned} \quad (4)$$

From (4) it can be seen that at the population level, the PH assumption no longer holds. If  $S_0(t|U=1)$  is left arbitrary, the model is the semiparametric Cox PH mixture cure model as described by Kuk and Chen [12], Peng and Dear [14] and Sy and Taylor [15].

2.2. Likelihood function

Suppose the data are of the form  $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \dots, n$ , where  $\delta_i$  is the censoring indicator with  $\delta_i = 1$  if  $t_i$  is uncensored and  $\delta_i = 0$  otherwise. It follows that, if  $\delta_i = 1$  then  $u_i = 1$ , and if  $\delta_i = 0$  then  $u_i$  is not observed, where  $u_i$  is the value taken by the random variable  $U_i$ . Under the assumption of type I censoring, the likelihood contribution for individual  $i$  is  $\pi_i(\mathbf{z}_i) f(t_i|u_i = 1, \mathbf{x}_i)$  for  $\delta_i = 1$  and  $1 - \pi_i(\mathbf{z}_i) + \pi_i(\mathbf{z}_i) S(t_i|u_i = 1, \mathbf{x}_i)$  for  $\delta_i = 0$ . The observed full likelihood is given by

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^n \{ \pi_i(\mathbf{z}_i) \lambda_0(t_i|u_i = 1) e^{\boldsymbol{\gamma}' \mathbf{x}_i} e^{-\Lambda_0(t_i|u_i = 1) \exp(\boldsymbol{\gamma}' \mathbf{x}_i)} \}^{\delta_i} \times \{ (1 - \pi_i(\mathbf{z}_i)) + \pi_i(\mathbf{z}_i) e^{-\Lambda_0(t_i|u_i = 1) \exp(\boldsymbol{\gamma}' \mathbf{x}_i)} \}^{(1 - \delta_i)} \tag{5}$$

where  $\Lambda_0(t|U=1) = \int_0^t \lambda_0(v|U=1) dv$  is the conditional cumulative hazard function.

When no cured fraction is assumed, i.e.  $\pi(\mathbf{z}_i) = 1$  for all  $\mathbf{z}_i$ , the likelihood function (5) reduces to the likelihood of the standard survival model.

3. THE PENALIZED LIKELIHOOD APPROACH

3.1. Penalized likelihood function

Most often the conditional hazard function can be expected to be smooth. A possible means to introduce such *a priori* knowledge is to penalize the likelihood by a term, which takes large value for rough functions. We choose the roughness penalty function to be the L2 norm of the second derivative of the conditional hazard function. The penalized likelihood function is then given by

$$pl(\boldsymbol{\gamma}, \boldsymbol{\beta}) = l(\boldsymbol{\gamma}, \boldsymbol{\beta}) - \kappa \int \lambda_0''(v|U=1)^2 dv \tag{6}$$

where  $l$  is the full log-likelihood given in (5) and  $\kappa > 0$  a positive smoothing parameter, which controls the trade-off between the fit to the data and the smoothness of the function. Maximization of (6) in the desired class of function defines the maximum penalized likelihood estimators (MPnLE)  $\hat{\lambda}_0(\cdot|U=1)$ ,  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\gamma}}$ .

3.2. Approximation by splines

The estimator  $\hat{\lambda}_0(\cdot|U=1)$  cannot be computed explicitly but can be approximated by a linear combination of  $m$  cubic M-splines, which are normalized B-splines. An M-spline of order  $k$  is defined as

$$M_j(x|k) = \begin{cases} \frac{k[(x - t_j)M_j(x|k-1) + (t_{j+k} - x)M_{j+1}(x|k-1)]}{(k-1)(t_{j+k} - t_j)}, & t_j \leq x < t_{j+k} \\ 0 & \text{elsewhere} \end{cases}$$

with

$$M_j(x|1) = \begin{cases} \frac{1}{(t_{j+1} - t_j)} & \text{if } t_j \leq x < t_{j+1} \\ 0 & \text{elsewhere} \end{cases}$$

where  $t_1, \dots, t_m$  is a sequence of increasing knots. Each  $M_j(x|k)$  is zero outside the interval  $[t_j, t_{j+k}]$  and hence is non-zero over  $k$  intervals, and over each interval there are  $k$  non-zero M-splines. We use splines of order 4 (also called cubic splines). To each spline, we associate an I-spline, i.e.

$$I_j(x|k) = \int_0^x M_j(u|k) du$$

Each  $M_j$  is a piecewise polynomial of degree  $k-1$  and each associated  $I_j$  is a piecewise polynomial of degree  $k$  defined as (for  $t_j \leq x < t_{j+1}$ )

$$I_h(x|k) = \begin{cases} 0 & \text{if } h > j \\ \sum_{l=h}^j (t_{l+k+1} - t_l) \frac{M_l(x|k+1)}{k+1} & \text{if } j-k+1 \leq h < j \\ 1 & \text{if } h < j-k+1 \end{cases}$$

Note that M-splines are non-negative and I-splines are monotonically increasing. The monotonicity constraint for a function represented on a basis of I-splines can thus be fulfilled by constraining the coefficients to be positive. The estimator  $\hat{\Lambda}_0(\cdot|U=1)$  can be approximated by a linear combination of  $m$  I-splines  $\tilde{\Lambda}_0(\cdot|U=1) = \sum_{j=1}^m g(\theta_j) I_j(\cdot)$  where  $g(\theta_j) \geq 0 \forall j$ . In practice we use  $g(\theta_j) = \theta_j^2$ . By differentiating, we get the conditional baseline hazard function with the same vector of coefficients  $\theta = (\theta_1, \dots, \theta_m)^T$ , such that  $\tilde{\lambda}_0(\cdot|U=1) = \sum_{j=1}^m g(\theta_j) M_j(\cdot)$ . The resulting function is a smooth positive function, defined on the positive real line and has continuous second derivatives. A spline function is completely defined by a sequence of increasing knots  $(t_1, \dots, t_l)$  and the coefficients  $\theta$ . The first knot is set at 0 and the last knot is set at the last (possibly censored) follow-up time point, so that the hazard function is defined on the whole follow-up period, including censored times greater than the last observed event time. Other knots can be either put equidistantly between the first and last one, or can be defined taking into consideration the observed times. Therefore, we have  $m = l + 2$  parameters to estimate the conditional hazard function. Theoretically, the more knots, the better the approximation. Increasing the number of knots does not deteriorate the MPnLE: this is because the degree of smoothing in the penalized likelihood method is tuned by the smoothing parameter  $\kappa$  and not by the number of splines. On the other hand, once a sufficient number of knots is established, there is no advantage in adding more. Moreover, the more knots, the longer the running time, especially if the method involves a search for the smoothing parameter. Furthermore, some numerical problems (slow convergence or failure to invert the Hessian matrix) can arise, particularly for a large number of knots. Generally, we found that a good choice for the number of knots is between 7 and 20.

### 3.3. Estimation procedure

We implemented the method with a Fortran program. The approximation  $\tilde{\lambda}_0(t|U=1)$  of  $\lambda_0(t|U=1)$  is the function belonging to the space generated by the basis of splines, which maximizes  $\text{pl}(\lambda)$ . The estimate  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$  for a fixed  $\kappa$  is obtained by maximizing the log-likelihood using the robust Marquardt algorithm [22], which is a combination of a Newton–Raphson algorithm and a steepest descent algorithm. This algorithm has the advantage of being more stable than the Newton–Raphson algorithm while preserving its fast convergence property near the maximum. Convergence is attained when the difference between two consecutive log-likelihoods is small, the coefficients are stable and the gradient is small enough. This maximization method performs well in most cases. To avoid the identifiability problem that may sometimes arise due to the absence of constraint on the conditional survival function, initial MPnLE values for  $\boldsymbol{\theta}$  and  $\boldsymbol{\gamma}$  may first be obtained by fitting a standard survival model to the uncensored subjects.

The variance of regression parameters are computed using the inverse of the matrix of the second derivatives of the penalized likelihood.

### 3.4. Selection of the smoothing parameter $\kappa$

Because we model the conditional hazard function and we do not constrain the conditional survival function at the right tail, parameters estimates may be sensitive to the choice of  $\kappa$ .

In our estimation method, an empirical estimate of  $\kappa$  can be provided or the smoothing parameter can be chosen by maximizing an approximate cross-validation score [23]

$$\text{LCV}_a = l(\cdot) - \text{Tr}(H_{\text{pl}}^{-1} H_l)$$

where  $l$  is the likelihood given in (5) with the corresponding Hessian matrix  $H_l$  and  $H_{\text{pl}}$  is the Hessian matrix based on the penalized likelihood [19]. In the presence of very late failure times, the automatic selection of the smoothing parameter may give MPnLE that suggest the absence of a cured fraction (i.e.  $\pi(\mathbf{z}) \approx 1$  and  $\hat{S}(t_{\max}|U=1) > 0$ ), where  $t_{\max}$  is the last follow-up times. However, the existence of very late failure times may also preclude the application of mixture cure models because it may indicate the absence of a cured fraction or insufficient follow-up.

## 4. SIMULATION STUDY

In this section we present the results from two simulation studies to illustrate the performance of the proposed estimation method.

The first simulation study aims at evaluating the performance of the model in a regression context. Data were generated from the logistic-Weibull mixture cure model. Specifically, the cured fraction is modeled through  $1 - \pi(z) = [1 + \exp(\beta_0 + \beta_1 z)]^{-1}$  and the latency part is given by  $S(t|U=1) = \exp(-(\zeta t)^\rho \exp(\gamma z))$ . The covariate  $z$  was fixed by design and binary, with half of the individuals having  $z=1$ . Censoring times  $C$  were derived from an exponential distribution with censoring rate  $\lambda_c$ . Each individual was followed for at most  $\tau_{\max} = 10$ , so that the data were  $(t, \delta, z)$  where  $t = \min(C, 10)$  when  $U_i = 0$  and  $t = \min(T, C, 10)$  when  $U_i = 1$ . Various values of  $\beta_0$ ,  $\beta_1$ ,  $\gamma$  and  $\lambda_c$  were considered.

Values for  $\beta_0$  were either  $-1.3863$ ,  $0$  or  $1.3863$ , so that the cured fraction was 80, 50 or 20 per cent in the baseline group ( $z=0$ ), respectively. Values for  $\beta_1$  were either  $0$  or  $-1$ , meaning

Table I. Simulation results: low censoring ( $\lambda_c=0.05$ ).

| Parameter  | True value | MPnLE estimates |           |           |               | Semiparametric estimates |           |
|--|------------|-----------------|-----------|-----------|---------------|--------------------------|-----------|
|  |            | Bias            | Mean var. | Emp. var. | CR (per cent) | Bias                     | Emp. var. |
| <i>Low cured fraction (20 per cent in the baseline group)</i>  |            |                 |           |           |               |                          |           |
| $\beta_0$  | 1.3863     | 0.0548          | 0.0839    | 0.0790    | 95.6          | 0.0407                   | 0.0814    |
| $\beta_1$  | -1         | -0.0073         | 0.1339    | 0.1191    | 95.8          | 0.0077                   | 0.1223    |
| $\gamma$   | 0.5        | 0.0162          | 0.0383    | 0.0395    | 95.6          | -0.0021                  | 0.0405    |
| <i>Mild cured fraction (50 per cent in the baseline group)</i> |            |                 |           |           |               |                          |           |
| $\beta_0$  | 0          | 0.0130          | 0.0486    | 0.0429    | 96.8          | 0.0681                   | 0.0444    |
| $\beta_1$  | -1         | -0.0096         | 0.1075    | 0.1007    | 96.4          | -0.0019                  | 0.1027    |
| $\gamma$   | 0.5        | 0.0587          | 0.0763    | 0.0962    | 95.2          | 0.0353                   | 0.0935    |
| <i>High cured fraction (80 per cent in the baseline group)</i> |            |                 |           |           |               |                          |           |
| $\beta_0$  | -1.3863    | -0.0371         | 0.0769    | 0.0748    | 95.4          | -0.0414                  | 0.0764    |
| $\beta_1$  | -1         | -0.0014         | 0.2397    | 0.2115    | 97.2          | -0.0051                  | 0.2407    |
| $\gamma$   | 0.5        | 0.0308          | 0.2521    | 0.2651    | 94.0          | 0.0435                   | 0.3482    |

that the covariate  $z$  has no effect on the incidence part or that it increases the cured fraction in individuals with  $z=1$ , respectively. In the same way, values for  $\gamma$  were either 0 or 0.5, meaning that the covariate  $z$  has no effect on the latency part or that it accelerates the occurrence of failure in susceptible individuals with  $z=1$ . We set  $\xi=0.3$  and  $\rho=2.2$ . The censoring rate  $\lambda_c$  was either 0.05, 0.1 or 0.3 representing, low, mild and heavy censoring, respectively (corresponding to an average 10, 25 and 50 per cent of censored susceptible individuals, respectively).

The penalized likelihood approach was applied to each sample with seven equidistantly distributed knots defining the hazard function. The smoothing parameter was chosen automatically for each sample using the cross-validation method explained previously.

For each configuration, 500 samples were generated, with sample size 200. For each sample, the value of  $\hat{S}_0(t_{\max}|U=1)$  was recorded, where  $t_{\max}$  is the last follow-up time.

Tables I–III present the estimated biases for regression parameters  $\beta_0$ ,  $\beta_1$  and  $\gamma$  based on 500 samples. To evaluate the adequacy of estimated variances, we compared the empirical variance, i.e. the variance of estimates, with the mean of estimated variances. The coverage rates of the normal approximation 95 per cent confidence interval, based on the point estimates and the estimated variances are also shown for each regression parameter. Results from the semiparametric mixture cure model of Peng and Dear [14] and Sy and Taylor [15] applied to the same samples with a zero-tail constraint are also presented in terms of biases and empirical variance. Results are only presented here for  $\beta_1=-1$  and  $\gamma_1=0.5$ , but conclusions were almost identical for other simulated values.

It can be seen from Tables I to III that point estimates have little biases in most cases and that the proposed estimation method performs as well as the semiparametric approach. The biases on  $\hat{\beta}_1$  and  $\hat{\gamma}$  are higher in case of heavy censoring, meaning that a reasonable proportion of non-susceptibles surviving for long enough is needed to ensure a good estimate of the incidence and latency effect of covariates. Also note that the penalized likelihood approach yields smaller biases for  $\hat{\beta}_0$  than the semiparametric modeling, due to the absence of tail constraint. In addition, the empirical variance and the mean of estimated variances agree pretty well and coverage rates are

Table II. Simulation results: mild censoring ( $\lambda_c=0.1$ ).

| Parameter  | True value | MPnLE estimates |           |           |               | Semiparametric estimates |           |
|--|------------|-----------------|-----------|-----------|---------------|--------------------------|-----------|
|  |            | Bias            | Mean var. | Emp. var. | CR (per cent) | Bias                     | Emp. var. |
| <i>Low cured fraction (20 per cent in the baseline group)</i>  |            |                 |           |           |               |                          |           |
| $\beta_0$  | 1.3863     | 0.0337          | 0.1053    | 0.1022    | 94.6          | 0.0099                   | 0.0987    |
| $\beta_1$  | -1         | 0.0967          | 0.1665    | 0.1490    | 94.6          | 0.1134                   | 0.1501    |
| $\gamma$   | 0.5        | -0.0205         | 0.0455    | 0.0456    | 95.6          | -0.0371                  | 0.0486    |
| <i>Mild cured fraction (50 per cent in the baseline group)</i> |            |                 |           |           |               |                          |           |
| $\beta_0$  | 0          | 0.0087          | 0.0589    | 0.0504    | 97.2          | -0.0034                  | 0.0513    |
| $\beta_1$  | -1         | -0.0501         | 0.1267    | 0.1111    | 97.6          | -0.0398                  | 0.1142    |
| $\gamma$   | 0.5        | 0.0760          | 0.0946    | 0.0962    | 95.2          | 0.0672                   | 0.1193    |
| <i>High cured fraction (80 per cent in the baseline group)</i> |            |                 |           |           |               |                          |           |
| $\beta_0$  | -1.3863    | -0.0413         | 0.0918    | 0.0908    | 96.0          | -0.0567                  | 0.0911    |
| $\beta_1$  | -1         | -0.0050         | 0.2771    | 0.2540    | 96.4          | -0.0023                  | 0.2940    |
| $\gamma$   | 0.5        | 0.0720          | 0.2933    | 0.3052    | 95.6          | 0.1116                   | 0.4482    |

Table III. Simulation results: heavy censoring ( $\lambda_c=0.3$ ).

| Parameter  | True value | MPnLE estimates |           |           |               | Semiparametric estimates |           |
|--|------------|-----------------|-----------|-----------|---------------|--------------------------|-----------|
|  |            | Bias            | Mean var. | Emp. var. | CR (per cent) | Bias                     | Emp. var. |
| <i>Low cured fraction (20 per cent in the baseline group)</i>  |            |                 |           |           |               |                          |           |
| $\beta_0$  | 1.3863     | 0.0322          | 0.1866    | 0.1783    | 97.6          | -0.1244                  | 0.1886    |
| $\beta_1$  | -1         | 0.1250          | 0.3079    | 0.2847    | 96.4          | 0.1822                   | 0.2933    |
| $\gamma$   | 0.5        | 0.0301          | 0.0904    | 0.0944    | 95.4          | 0.0099                   | 0.0964    |
| <i>Mild cured fraction (50 per cent in the baseline group)</i> |            |                 |           |           |               |                          |           |
| $\beta_0$  | 0          | 0.0299          | 0.1415    | 0.1473    | 97.4          | -0.1470                  | 0.1937    |
| $\beta_1$  | -1         | -0.1591         | 0.2474    | 0.2594    | 95.2          | -0.1624                  | 0.2194    |
| $\gamma$   | 0.5        | 0.0529          | 0.2269    | 0.2269    | 92.4          | 0.0460                   | 0.2608    |
| <i>High cured fraction (80 per cent in the baseline group)</i> |            |                 |           |           |               |                          |           |
| $\beta_0$  | -1.3863    | -0.0372         | 0.1813    | 0.1776    | 96.0          | -0.1707                  | 0.1683    |
| $\beta_1$  | -1         | -0.1788         | 0.3952    | 0.3845    | 97.8          | -0.1452                  | 0.4521    |
| $\gamma$   | 0.5        | 0.1723          | 0.5055    | 0.5194    | 90.7          | 0.2023                   | 0.6507    |

good in most cases with mild censoring, suggesting that the variance estimators perform quite well. In case of heavy censoring and small proportions of susceptible individuals ( $\beta_0 = -1.3863$ ), the estimate  $\hat{\gamma}$  is not stable, due to an insufficient number of observed events, which is reflected by high observed and estimated variances. However, the variances estimators still perform well and coverage rates are still reasonable.

In the second simulation study, we focused on the estimation of the baseline conditional hazard function. To illustrate the flexibility of the conditional hazard function estimator based on M-splines, we generated a mixture of two Weibull hazard functions, such that  $S(t|U=1) = pS_1(t|U=1) + (1-p)S_2(t|U=1)$  where  $S_1(t|U=1) = \exp(-(\xi_1 t)^{\rho_1} \exp(\gamma Z))$  and  $S_2(t|U=1) = \exp(-(\xi_2 t)^{\rho_2} \exp(\gamma Z))$ . We set  $\xi_1 = 1.0$ ,  $\rho_1 = 1.5$ ,  $\xi_2 = 0.1$ ,  $\rho_2 = 5.0$ ,  $\gamma = 0.5$  and  $p = 0.4$ .



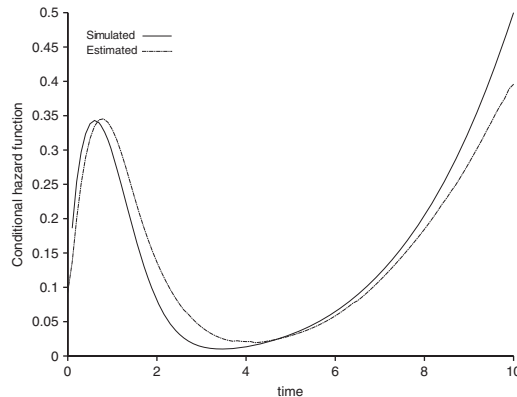


Figure 1. The simulated conditional hazard functions in simulation 2 and the mean estimated one (over 500 simulated sets).

The resulting baseline conditional hazard function is shown in Figure 1. The censoring times were generated using the uniform distribution  $U[0; 20]$ . A logistic model was used to model the cured fraction such that  $1 - \pi(z) = [1 + \exp(\beta_0 + \beta_1 z)]^{-1}$ , with  $\beta_0 = 0$  and  $\beta_1 = -0.5$ .

Under this setting 500 random samples were generated with sample size 1000, 400 and 200. The penalized likelihood approach was applied with an automatic search for the smoothing parameter with 14 equidistantly distributed knots. The performances of the estimates of the conditional hazard and survival functions were evaluated in term of biases at different time points, i.e. 0.6, 3.5, 6.5 and 10. The variance of  $\hat{\lambda}_0(t|U=1)$  was computed using the delta method, and we compared the empirical variance, i.e. the variance of estimates, with the mean of estimated variances. The results are shown in Table IV. With sample size 1000, biases are small at all time points and suggest good properties of the estimators. Biases increase with smaller sample sizes. Empirical variances and the means of estimated variances agree in most cases and coverage rates are satisfactory. The largest variances are observed at time point 10, when the number in the set at risk is reduced. The good fit between the simulated conditional baseline hazard function and the mean estimated curves are illustrated graphically on Figure 1.

## 5. APPLICATION

To illustrate the penalized likelihood approach, we consider the data from a study of patients with tonsil cancer who received radiation therapy [24]: 672 patients with squamous cell carcinoma of the tonsil were collected by the investigators. The data have already been explored using a semiparametric PH mixture model [15] and a semiparametric accelerated mixture cure model [25]. In this example, local recurrence is considered as the event and failure time is time from initial treatment to local recurrence. After an approximate 15 years follow-up period, 206 events were observed. The Kaplan–Meier plot of the marginal survival function levels off with a large and stable plateau at the right tail, ensuring the appropriateness of the mixture cure model approach (not shown here). Six covariates are considered here, including T-stage, which has four levels

Table IV. Simulation results: bias, empirical variance ( $\times 100$ ), mean estimated variance ( $\times 100$ ) and coverage rate for the estimation of the conditional hazard function.

| Sample size | Time point | True value | Absolute bias | Mean variance ( $\times 100$ ) | Empirical variance ( $\times 100$ ) | Coverage rate (per cent) |
|-------------|------------|------------|---------------|--------------------------------|-------------------------------------|--------------------------|
| $n = 1000$  | 0.6        | 0.3430     | -0.0086       | 0.1419                         | 0.1655                              | 93.5                     |
|             | 3.5        | 0.0102     | 0.0096        | 0.0062                         | 0.0051                              | 95.8                     |
|             | 6.5        | 0.0892     | -0.0080       | 0.0275                         | 0.0277                              | 95.6                     |
|             | 10.0       | 0.5000     | -0.0767       | 0.5604                         | 0.4595                              | 97.5                     |
| $n = 400$   | 0.6        | 0.3430     | -0.0592       | 0.2399                         | 0.2925                              | 92.5                     |
|             | 3.5        | 0.0102     | 0.0108        | 0.0119                         | 0.0087                              | 92.6                     |
|             | 6.5        | 0.0882     | -0.0090       | 0.0590                         | 0.0482                              | 95.5                     |
|             | 10.0       | 0.5000     | -0.0906       | 1.0158                         | 0.7965                              | 98.7                     |
| $n = 200$   | 0.6        | 0.3430     | -0.0959       | 0.3642                         | 0.4140                              | 90.4                     |
|             | 3.5        | 0.0102     | 0.0152        | 0.0324                         | 0.0238                              | 97.5                     |
|             | 6.5        | 0.0882     | -0.0146       | 0.0992                         | 0.0883                              | 96.2                     |
|             | 10.0       | 0.5000     | -0.0830       | 1.7160                         | 1.3645                              | 98.8                     |

( $T_1; T_2; T_3; T_4$ ), neck node status, which has two levels (negative (N0) and positive (N+)), sex, and the continuous variables age, total dose and overall treatment duration.

In our application of the penalized likelihood approach, we used 14 knots equidistantly placed between 0 and the last follow-up time (14.46 years). Adding more knots did not improve the likelihood significantly and did not result in modifications of the baseline conditional hazard function (not shown). The results were not sensitive to the location of knots. The smoothing parameter was chosen automatically based on the cross-validation criterion and was almost similar for 12 knots or more (about  $10^{-5}$ ). Standard errors for the parameter estimates were computed using the inverse of the second derivative of the penalized likelihood.

Table V gives the results from the proposed penalized likelihood estimation method and from the Cox PH mixture cure model of Sy and Taylor [14]. Parameter estimates are very similar between the two estimation methods and variance estimates are in close agreement. T-stage acts significantly on the incidence part, patients with high T-stage being cured less often. The effect of T-stage is also significant on latency ( $p = 0.003$ ), with patients with T-stage = 2 having delayed recurrences. Node status is found to be significant in the latency part ( $p = 0.012$ ) but not in the incidence part ( $p = 0.087$ ), with earlier recurrence times for those with positive nodes. The total dose and treatment duration are highly significant for incidence but not for latency. Higher doses reduce the recurrence rate while patients with long treatment duration are at high risk to fail. Sex is not significant on either incidence or latency. While it is not significant for the incidence part, the effect of age is significant on latency, local recurrence being delayed in older patients.

Figures 2 and 3 illustrate the effect of age and T-stage on incidence and latency. The estimated conditional and marginal hazard functions and the corresponding survival functions based on the penalized likelihood approach were computed for a male, node positive patient who received 64 Gy total dose and had overall treatment duration of 45 days. The number at-risk in the susceptible group, based on the model estimates, was 326.0 at  $t = 0$ , 158.43 at 6 months, 45.66 at 2 years, 11.20 at four years and only 1.85 after 8 years of follow-up. The effect of age (Figure 1) is computed for fixed T-stage = 2. The opposite effect of age on incidence and latency is illustrated by survival and

Table V. Estimates from the penalized likelihood model mixture cure model (MPnLE) and the semiparametric PH mixture cure model.

|                        | MPnLE estimates |       |                    | Semiparametric PH mixture model |       |                    |
|------------------------|-----------------|-------|--------------------|---------------------------------|-------|--------------------|
|                        | Estimate        | SE    | ( <i>p</i> -Value) | Estimate                        | SE    | ( <i>p</i> -Value) |
| <i>Incidence model</i> |                 |       |                    |                                 |       |                    |
| Intercept              | -0.122          | 0.999 |                    | -0.181                          | 1.030 |                    |
| T-stage                |                 |       | 0.013*             |                                 |       |                    |
| T <sub>2</sub>         | 0.871           | 0.354 | (0.014)            | 0.852                           | 0.357 | (0.017)            |
| T <sub>3</sub>         | 1.608           | 0.342 | (<0.001)           | 1.655                           | 0.345 | (<0.001)           |
| T <sub>4</sub>         | 2.192           | 0.426 | (<0.001)           | 2.198                           | 0.430 | (<0.001)           |
| Node                   | 0.304           | 0.199 | (0.087)            | 0.355                           | 0.204 | (0.082)            |
| Total dose (Gy)        | -0.079          | 0.018 | (<0.001)           | -0.077                          | 0.018 | (<0.001)           |
| Duration (per 10 days) | 0.462           | 0.124 | (<0.001)           | 0.463                           | 0.127 | (<0.001)           |
| Sex: male              | 0.043           | 0.211 | (0.836)            | 0.116                           | 0.215 | (0.590)            |
| Age (per 10 years)     | 0.109           | 0.093 | (0.243)            | 0.134                           | 0.097 | (0.167)            |
| <i>Latency model</i>   |                 |       |                    |                                 |       |                    |
| T-stage                |                 |       | 0.003*             |                                 |       |                    |
| T <sub>2</sub>         | -0.707          | 0.358 | (0.048)            | -0.625                          | 0.365 | (0.087)            |
| T <sub>3</sub>         | -0.041          | 0.341 | (0.904)            | -0.108                          | 0.352 | (0.759)            |
| T <sub>4</sub>         | 0.415           | 0.378 | (0.272)            | 0.385                           | 0.383 | (0.315)            |
| Node                   | 0.456           | 0.182 | (0.012)            | 0.339                           | 0.188 | (0.071)            |
| Total dose (Gy)        | -0.005          | 0.012 | (0.673)            | -0.005                          | 0.014 | (0.721)            |
| Duration (per 10 days) | -0.031          | 0.074 | (0.675)            | -0.007                          | 0.078 | (0.920)            |
| Sex: male              | 0.190           | 0.188 | (0.314)            | 0.065                           | 0.198 | (0.743)            |
| Age (per 10 years)     | -0.253          | 0.096 | (0.009)            | -0.303                          | 0.097 | (0.002)            |

\*Wald  $\chi^2$  with 3 df.

hazard curves that cross. The effect of T-stage is computed for fixed age=60 years. Similarly the marginal hazard curves cross after 1 year, illustrating the different effects of T-stage on incidence and latency.

Because of a very long follow-up period beyond the last failure time, which ensures the appropriateness of the zero-tail constraint, the estimated conditional and marginal survival curves from the Cox PH mixture cure model (not shown here) are similar (although not smooth) to those obtained with our approach.

The conditional hazard curves for both age and T-stage show a bimodal shape between 0 and 4 years, with hazard peaking first around 0.5 year and then around 2.5 years. This was not modified by changing the location of knots. The hazard function is not plotted beyond 4 years, because of the very small number in the risk set beyond this date, but goes to infinity after the last event time (not shown). Note that this bimodal shape could not be identified using a Weibull mixture cure mixture model, and illustrates the flexibility of the approximation of the hazard function using splines. The reason for the two modes in the hazard is unclear. It may be due to scheduling of the patient's visits to the doctor's clinic, at which time recurrences may be first detected. First, the investigators of the study excluded cases with less than one month disease-free follow-up [24]: this explains why the estimate of the hazard function starts from zero. Second, it is possible (although not described in [24]) that the recurrence has been assessed at visit times with similar schedules

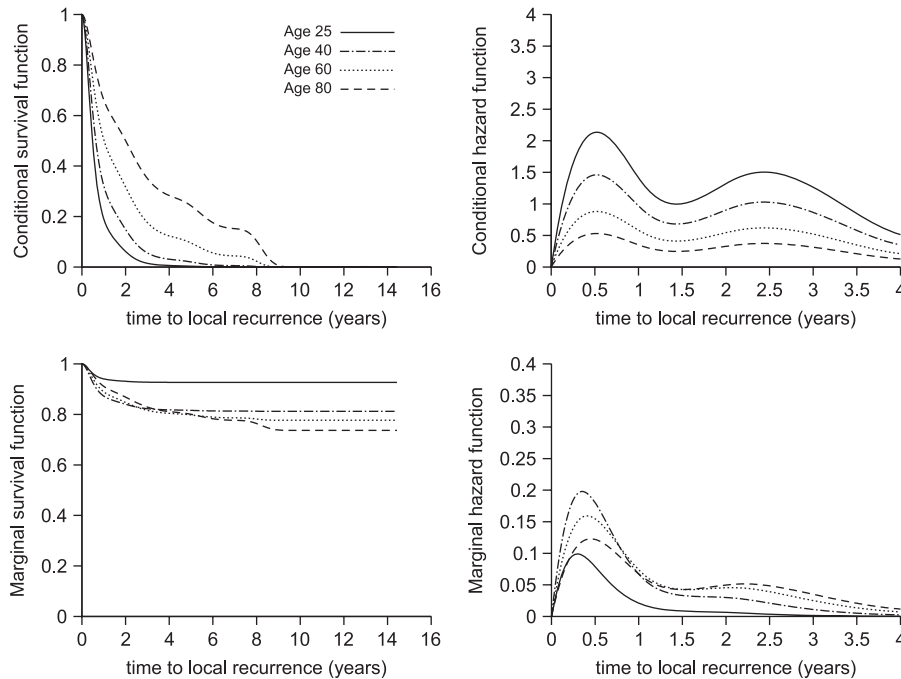


Figure 2. Estimated conditional and marginal survival (left) and hazard (right) functions from the penalized likelihood approach for four different ages. For each age, the other covariates are fixed at selected values: sex = male, total dose = 64 Gy, overall treatment duration = 45 days, node positive, T-stage = 2.

for most of the patients. This could have been treated by methods allowing for interval censoring if the visit times had been available.

To compare the fit provided by the two-component mixture cure model against the standard survival model with no cured fraction, we used the normalized likelihood-based cross validation criterion (LCV), which is a nearly unbiased estimator of the Kullback–Leibler criterion [23] and reduces to the normalized Akaike information criterion for parametric models (without penalized likelihood). According to Commenges *et al.* [23], a difference less than  $10^{-3}$  in the LCV values between two models is considered as ‘negligible’ (i.e. the fit provided by the two models are comparable), whereas a difference greater than  $10^{-2}$  is considered as ‘moderate’ to ‘large’. Under the penalized likelihood approach, the results were  $LVC = 1.5070$  for the one component (standard) survival model and  $LCV = 1.4605$  for the two-component mixture cure model indicating that the introduction of a cured fraction provides a better fit to the data. This conclusion was not influenced by the number and location of knots. We also used the LCV to compare the fit provided by the proposed approach and a parametric Weibull approach including the standard Weibull survival model and the logistic-Weibull mixture cure model. The results were  $LCV = 1.5017$  for the Weibull mixture cure model and  $LCV = 1.7547$  for the standard Weibull model. Here again the introduction of a cured fraction improves the fit to the data. Moreover, it can be concluded from the comparison of the LCV values that the penalized likelihood approach is preferred to the Weibull mixture cure model.

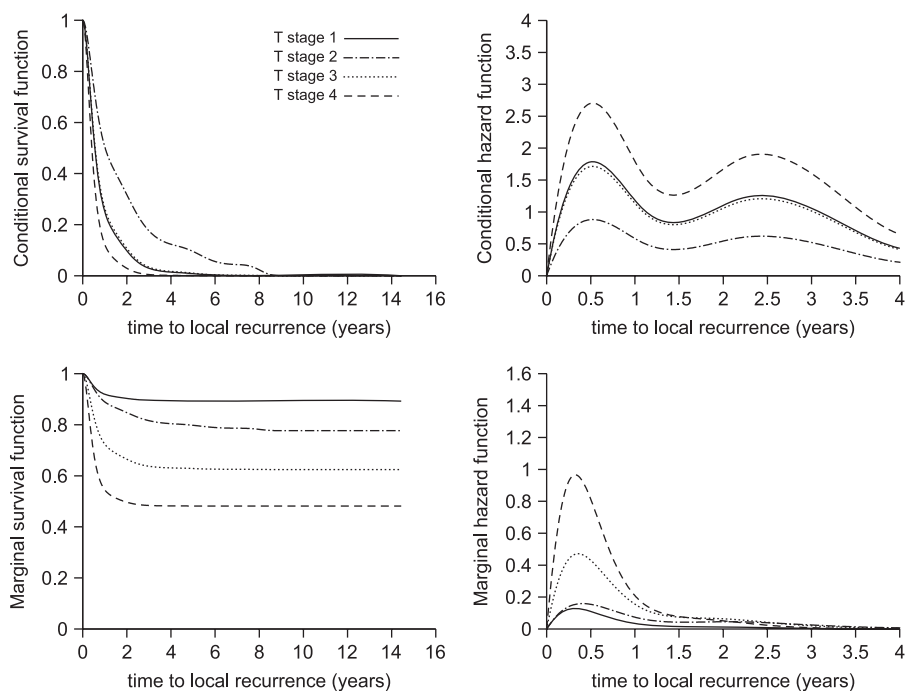


Figure 3. Estimated conditional and marginal survival (left) and hazard (right) functions from the penalized likelihood approach for four different T-stages. For each T-stage, the other covariates are fixed at selected values: sex= male, total dose=64 Gy, overall treatment duration=45 days, node positive, age=60 years.

## 6. DISCUSSION

In this paper we propose a new estimation method applied to mixture cure models. Our work was mainly motivated by two major issues regarding estimation procedures in this kind of model: the estimation of the conditional survival/hazard function, and the closely related estimation of the variance of regression parameters. The proposed approximation of the conditional hazard function using splines yields a smooth and much more flexible estimate than standard fully parametric distributions. Moreover, estimation could be performed by direct maximization of the full-likelihood function, whereas the EM-algorithm is mostly used in semiparametric mixture cure models [14–16, 25, 26]. Although this estimation procedure works well without specifying any tail constraint, we did not completely rule out the identifiability issues commonly met with mixture cure models. The choice of the smoothing parameter  $\kappa$  based on the approximate cross-validation score may lead to an improper conditional survival estimate in the presence of late failure time. This issue corresponds, however, to cases where the application of mixture cure models may not be justified due to insufficient follow-up or because the presence of a cured fraction is questionable.

Alternatively, data can be analyzed utilizing statistical models that account for heterogeneity among individuals. These models, also known as frailty models, differ from cure models in that they assume all individuals eventually experience the event of interest with varying risks that are greater

than zero [27, 28]. The proportion of individuals considered to be cured in the former models are generally considered as having a low risk of experiencing the event in the latter models. As noted by Sy and Taylor [15], the mixture cure model is a special case of a multiplicative frailty model, in which the hazard for an individual, conditional on  $U$ , can be written as  $\lambda(t|U, \mathbf{x}) = U\lambda(t|U=1, \mathbf{x})$ . As a frailty variable,  $U$  is not entirely observable since an individual becomes labeled as  $U=1$  if an event is observed. Usually frailties are assumed to follow a distribution such as the gamma, inverse Gaussian or positive stable distribution. These frailty distributions do not allow individuals to have zero risk (i.e.  $\Pr(U=0)=0$ ), and the standard frailty models do not allow a cured proportion. As an extension of the parametric family, Aalen [29] considered a compound Poisson distribution, which allows a positive probability for the risk to be zero. A comparison of parametric mixture cure models, gamma frailty mixture models and compound Poisson models was performed by Price and Manutunga [30] in a one-sample, no-covariate analysis of a leukemia data set. Their results suggest that gamma frailty mixture models and compound Poisson models could perform better than parametric mixture cure models. However, these models only account for heterogeneity among individuals in the latency part, not in the incidence part. Parametric mixture cure models with random effects have recently been proposed [31], but the choice of the frailty distribution and of their covariance matrix may be an important issue. Further work is needed to address these issues using a penalized likelihood approach.

#### ACKNOWLEDGEMENTS

The authors thank the referees for their helpful comments and suggestions. These led to considerable improvements in the paper.

#### REFERENCES

1. Jones DR, Powles RL, Machin D, Sylvester RJ. On estimating the proportion of cured patients in clinical studies. *Biometrie-Praximetrie* 1981; **21**:1–11.
2. Goldman AI. Survivorship analysis when cure is a possibility: a Monte Carlo study. *Statistics in Medicine* 1984; **3**:153–163.
3. Ghitany ME, Maller RA, Zhou S. Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis* 1994; **49**:218–241.
4. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B* 1949; **11**:15–44.
5. Gamel JW, Vogel RL. Comparison of parametric and non-parametric survival methods using simulated clinical data. *Statistics in Medicine* 1997; **16**:1629–1643.
6. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982; **38**:1041–1046.
7. Gordon NH. Application of the theory of finite mixtures for the estimation of ‘cure’ rates of treated cancer patients. *Statistics in Medicine* 1990; **9**:397–407.
8. Cantor AB, Shuster JJ. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine* 1992; **11**:931–937.
9. Shao Q, Zhou X. A new parametric model for survival data with long-terms survivors. *Statistics in Medicine* 2004; **23**:3525–3543.
10. Peng Y, Dear KB, Denham JW. A generalized F mixture model for cure rate estimation. *Statistics in Medicine* 1998; **17**:813–830.
11. Cox DR. Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
12. Kuk AYC, Chen C. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* 1992; **79**:531–541.

13. Taylor JMG. Semi-parametric estimation in failure time mixture models. *Biometrics* 1995; **51**:899–907.
14. Peng Y, Dear KB. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; **56**:237–243.
15. Sy JP, Taylor JMG. Standard errors for the Cox proportional hazards cure model. *Mathematical and Computer Modelling* 2001; **33**:1237–1251.
16. Sy JP, Taylor JMG. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; **56**:227–236.
17. Peng Y. Estimating baseline distribution in proportional hazards cure models. *Computational Statistics and Data Analysis* 2003; **42**:187–201.
18. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996; **11**:89–121.
19. Joly P, Commenges D, Letenneur L. A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics* 1998; **54**:185–194.
20. Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002; **3**:433–443.
21. Lam KF, Fong DY, Tang OY. Estimating the proportion of cured patients in a censored sample. *Statistics in Medicine* 2005; **24**:1865–1879.
22. Marquardt D. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics* 1963; **11**:431–441.
23. Commenges D, Joly P, Gégout-Petit A, Liqueur B. Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics* 2007; **34**:33–54.
24. Withers HR, Peters LJ, Taylor JMG, Owen JB, Morrison WH, Schultheiss TE, Keane T, O Sullivan B, van Dyk J, Gupta N, Wang CC, Jones CU, Doppke KP, Myint S, Thompson M, Parsons JT, Mendenhall WM, Dische S, Aird EGA, Henk JM, Bidmead MAM, Svoboda V, Chan Y, Hanlon AL, Peters TL, Hanks GE. Local control of carcinoma of the tonsil by radiation therapy: an analysis of patterns of fractionation in nine institutions. *International Journal of Radiation Oncology, Biology, Physics* 1995; **33**:549–562.
25. Li CS, Taylor JMG. A semi-parametric accelerated failure time cure model. *Statistics in Medicine* 2002; **21**:3235–3247.
26. Zhang J, Peng Y. A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine* 2006; **26**:3157–3171.
27. Aalen OO. Heterogeneity in survival analysis. *Statistics in Medicine* 1988; **7**:1121–1137.
28. Hougaard P, Myllegard P, Borch-Johnsen K. Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. *Biometrics* 1994; **50**:1178–1188.
29. Aalen OO. Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability* 1988; **2**:951–972.
30. Price DL, Manatunga AK. Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine* 2001; **20**:1515–1527.
31. Yau KKW, Ng ASK. Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma. *Statistics in Medicine* 2001; **20**:1591–1607.