COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 85 (2007) 173–180

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A SAS macro for parametric and semiparametric mixture cure models

Fabien Corbière, Pierre Joly*

EMI E0338 Biostatistique, Institut de Santé Publique et Développement, Université Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France

ARTICLE INFO

Article history:
Received 7 March 2006
Received in revised form
31 October 2006
Accepted 31 October 2006

Keywords:
Survival analysis
Cure models
EM algorithm
Parametric and semiparametric models
SAS macro

ABSTRACT

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. The aim of this paper is to propose a SAS macro to estimate parametric and semiparametric mixture cure models with covariates. The cure fraction can be modelled by various binary regression models. Parametric and semiparametric models can be used to model the survival of uncured individuals. The maximization of the likelihood function is performed using SAS PROC NLMIXED for parametric models and through an EM algorithm for the Cox's proportional hazards mixture cure model. Indications and limitations of the proposed macro are discussed and an example in the field of cancer clinical trials is shown.

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. In recent years, there has been a increasing interest in modelling survival data with long term survivors. Such data may arise from clinical trials, in which, even after an extended follow-up, no further events of interest are observed. Some people in the population may be considered as cured or non-susceptible (cured). Failing to account for such cured subjects would lead to incorrect inferences. Moreover, researchers may be interested in estimating the cured fraction.

Mixture cure models assume that the studied population is a mixture of susceptible (uncured) individuals, that may experience the event of interest, and non-susceptible (cured) individuals, that will never experience it [1]. This approach allows to estimate simultaneously whether the event of interest will occur, which is called incidence, and when it will occur, given that it can occur, which is called latency.

Let U be the indicator denoting 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, defined only when $U = 1$. The mixture cure model is given by

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

where $S(t|\mathbf{x}, \mathbf{z})$ is the unconditional survival function of T for the entire population, $S(t|U = 1, \mathbf{x}) = P(T > t|U = 1, \mathbf{x})$ the

* *Corresponding author*. Tel.: +33 5 57 57 45 79; fax: +33 5 56 24 00 81.
E-mail address: fabien.corbiere@isped.u-bordeaux2.fr (F. Corbière).
0169-2607/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2006.10.008

survival function for susceptible individuals given a covariate vector $\mathbf{x} = (x_1, \ldots, x_p)'$, and $\pi(\mathbf{z}) = P(U = 1|\mathbf{z})$ is the probability of being susceptible given a covariate vector $\mathbf{z} = (z_1, \ldots, z_q)'$, which may include the same covariates as $\mathbf{x}$. The survival function of cured patients can be set to one for all finite values of t because they will never experience the event of interest. Note that $S(t|\mathbf{x}, \mathbf{z}) \to 1 - \pi(\mathbf{z})$ as $t \to \infty$. When $\pi(\mathbf{z}_i) = 1$ for all $\mathbf{z}_i$, i.e. when there is no cured fraction, the mixture cure model reduces to the standard survival model. Various parametric and semiparametric specifications of $S(t|U = 1)$ have been proposed, leading to parametric and semiparametric mixture cure models [2].

In the next section, a brief description of parametric and semiparametric mixture cure models is presented as well as computational methodology. The macro and its requirements are described in Section 3 and a simulation study is shown in Section 4. In Section 5 an illustrative example is provided.

2. Computational methods and theory

2.1. *Parametric and semiparametric mixture cure models*

The effect of $\mathbf{z}$ on the probability of $\pi(\mathbf{z})$ can be modelled by the use of binary regression models, with logit link

$$\text{logit}(\pi(\mathbf{z})) = \beta_0 + \beta_1 z_1 + \cdots + \beta_q z_q = \boldsymbol{\beta}'\mathbf{z}$$

where β_0 is the intercept and $\boldsymbol{\beta}$ is the vector of regression parameters associated to $\mathbf{z}$. Other regression models include the probit link

$$\Phi^{-1}(\pi(\mathbf{z})) = \boldsymbol{\beta}'\mathbf{z}$$

where Φ is the distribution function of a standard normal distribution, and the complementary log–log link

$$\log(-\log(1 - \pi(\mathbf{z}))) = \boldsymbol{\beta}'\mathbf{z}$$

The conditional latency distribution $S(t|U = 1)$ can take the form of parametric or semiparametric distributions. Among the parametric models, exponential (EXP), Weibull (WB), lognormal (LN) and loglogistic (LG) are commonly used to model survival data. After reparametrization [3], these distributions can be expressed as

$$S(t|U=1) = \begin{cases} \exp[-\exp(\log t - \mu)], & \text{exponential;} \\ \exp\left[-\exp\left(\dfrac{\log t - \mu}{\sigma}\right)\right], & \text{Weibull;} \\ 1 - \Phi\left(\dfrac{\log t - \mu}{\sigma}\right), & \text{lognormal;} \\ \left[1 + \exp\left(\dfrac{\log t - \mu}{\sigma}\right)\right]^{-1}, & \text{loglogistic;} \end{cases} \quad (2)$$

Covariates can be included by parameterizing μ such as $\mu = \boldsymbol{\gamma}'\mathbf{x}$, where $\boldsymbol{\gamma}$ represents the vector of unknown regression parameters. These models are also known as parametric accelerated failure time (AFT) mixture cure models [4]. Since $\mathbf{x}$ acts multiplicatively on the scale parameter μ, it accelerates or decelerates the failure time of susceptible individuals.

In proportional hazards (PH) models, the conditional distribution of T is modelled by

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

where $S_0(t|U = 1)$ and $\lambda_0(t|U = 1)$ are the baseline conditional survival and hazard functions, respectively. The conditional cumulative hazard function is $\Lambda(t|U = 1) = \Lambda_0(t|U = 1)\exp(\boldsymbol{\gamma}'\mathbf{x})$, where $\Lambda_0(t|U = 1) = \int_0^t \lambda_0(v|U = 1)\,dv$. If $S_0(t|U = 1)$ is left arbitrary, the model is defined as the Cox's proportional hazards mixture cure model [5]. Note that the Weibull (and exponential) models are both AFT and PH models.

Through the vectors of regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, parametric and semiparametric mixture cure models are able to separate the covariate effects on the incidence and the latency.

2.2. *Likelihood*

Suppose the data are of the form $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, where δ_i is the censoring indicator with $\delta_i = 1$ if t_i is uncensored and $\delta_i = 0$ otherwise. The likelihood contribution for individual i is $\pi_i(\mathbf{z}_i)f(t_i|U = 1, \mathbf{x}_i)$ for $\delta_i = 1$ and $(1 - \pi_i(\mathbf{z}_i)) + \pi_i(\mathbf{z}_i)S(t_i|U = 1, \mathbf{x}_i)$ for $\delta_i = 0$, where $f(\cdot) = S(\cdot)\lambda(\cdot)$ is the conditional probability density function of T. The observed full likelihood is then given by

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \{\pi_i(\mathbf{z}_i)f(t_i|U=1,\mathbf{x}_i)\}^{\delta_i} \times \{(1 - \pi_i(\mathbf{z}_i)) + \pi_i(\mathbf{z}_i)S(t_i|U=1,\mathbf{x}_i)\}^{1-\delta_i} \quad (4)$$

When no cured fraction is assumed, i.e. $\pi(\mathbf{z}_i) = 1$ for all $\mathbf{z}_i$, the likelihood function (4) reduces to the likelihood of the standard survival model.

2.3. *Estimation procedures*

This section presents a brief description of the procedures used to estimate the parameters by maximizing the likelihood (4). A more detailed presentation can be found in Peng and Dear [6] and Sy and Taylor [7].

2.3.1. *Maximization of the likelihood function*

To estimate $\boldsymbol{\gamma}$, we must specify the failure time distribution of uncured subjects. For parametric mixture cure models, $f(\cdot|U = 1)$ and $S(\cdot|U = 1)$ can be defined by a few unknown parameters in (4). Therefore, maximum likelihood estimates are obtained via usual optimization methods as the Newton–Raphson method (PROC NLMIXED [8]). Asymptotic standard errors are obtained by inverting the Fisher's information matrix of second order derivatives of $\log(L)$.

Unlike in the standard Cox's proportional hazard, where little information is lost by eliminating $S_0(t)$, one cannot eliminate $S_0(t|U = 1)$ in the Cox' PH mixture cure model without losing information about $\boldsymbol{\beta}$. The EM algorithm provides a simple and efficient way to estimate separately $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $S_0(t|U = 1, \mathbf{x}_i)$.

From the introduction, U is the random variable denoting an individual is susceptible ($U = 1$) or non-susceptible ($U = 0$). 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. Given the u_i's, the complete-data full log–likelihood (4) is the sum of two independent components, l_I, which depends only on $\boldsymbol{\beta}$, and l_S, which depends only on $\boldsymbol{\gamma}$ and Λ_0 where

$$l_I(\boldsymbol{\beta}; \boldsymbol{u}) = \log \prod_{i=1}^{n} \pi(\mathbf{z}_i)^{u_i} (1 - \pi(\mathbf{z}_i))^{1-u_i} \tag{5}$$

and

$$l_S(\boldsymbol{\gamma}, \Lambda_0; \boldsymbol{u}) = \log \prod_{i=1}^{n} \lambda(t|U = 1, \mathbf{x}_i)^{\delta_i u_i} S(t|U = 1, \mathbf{x}_i)^{u_i} \tag{6}$$

with $\boldsymbol{u}$ the vector of u_i values. The EM algorithm starts with initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ and $S_0^{(0)}(t|U = 1)$. The E step in the (r)th iteration calculates the expectation of the complete log–likelihood function with respect to $\boldsymbol{u}$, conditional on the observed data and $\boldsymbol{\beta}^{(r)}$, $\boldsymbol{\gamma}^{(r)}$ and $S_0^{(r)}(t|U = 1)$, the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $S_0(t|U = 1)$ at the rth iteration. This is given by the following conditional expectation

$$\begin{aligned} y_i^{(r)} &= E\{u_i|\boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, S_0^{(r)}(t|U = 1)\} \\ &= \delta_i + (1 - \delta_i)\frac{\pi^{(r)}(\mathbf{z}_i)S^{(r)}(t|U = 1, \mathbf{x}_i)}{1 - \pi^{(r)}(\mathbf{z}_i) + \pi^{(r)}(\mathbf{z}_i)S^{(r)}(t|U = 1, \mathbf{x}_i)} \end{aligned} \tag{7}$$

which is the rth estimator of the probability of the ith individual being susceptible. Given $y_i^{(r)}$ the M step in the $(r + 1)$th iteration maximizes the expected complete log–likelihood function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to obtain $\boldsymbol{\beta}^{(r+1)}$, $\boldsymbol{\gamma}^{(r+1)}$ and $S_0^{(r+1)}(t|U = 1)$. The algorithm iterates until convergence on estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $S_0(t|U = 1)$.

Peng and Dear [6] and Sy and Taylor [7] proposed a Cox's partial-likelihood-type method to estimate $\boldsymbol{\gamma}$ semiparametrically without specifying $\lambda_0(t|U = 1)$. For this method (6) is approximated by

$$\sum_{j=1}^{k} \left[\boldsymbol{\gamma}' \mathbf{s}_{(j)} - d_j \log \left(\sum_{i \in R_j} u_i \exp(\boldsymbol{\gamma}' \mathbf{x}_i) \right) \right] \tag{8}$$

where k is the number of distinct uncensored failures times, d_j the number of uncensored observations at t_j, $\mathbf{s}_{(j)}$ the sum of covariate vectors associated with the uncensored observations at t_j and R_j is the risk set at t_j.

Note that (5) is the log–likelihood function of a binary regression model, and that (6) is similarly the log–likelihood function of the standard Cox's PH model, with the addition of the offset variable $\log(u_i)$. Therefore, the M step of the EM algorithm is equivalent to the separated maximization of l_I and l_S with standard regression models for binary variables (PROC LOGISTIC) and failure time data (PROC PHREG), respectively.

2.3.2. *Estimation of the conditional baseline survival function in semiparametric models*

The estimation of $S_0(t|U = 1)$ is of concern here, because it is needed in (7). Two non-parametric methods are discussed in Peng and Dear [6] and Sy and Taylor [7]. The first one is based on a profile likelihood estimate of $\Lambda_0(t|U = 1)$, similar to the Breslow's likelihood for the standard Cox's PH model [9]. The second one is derived from the product limit estimator (PLE) after Kalbfleisch and Prentice [10].

In order to obtain a good estimation for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ it is important for $\hat{S}_0(t|U = 1)$ to approach 0 as $t \to \infty$. However, the estimates from the Breslow or the PLE methods do not approach zero as $t \to \infty$ when there are censored survival times after t_k, where t_k is the last observed failure time. Setting $\hat{S}_0(t|U = 1) = 0$ for all $t \geq t_k$ allows for a proper distribution function for susceptible individuals and avoid identifiability problems [11,12]. However, this zero-tail constraint implies that individuals with survival times greater than t_k are all considered as non-susceptible or cured, which may appear to be a strong assumption. Peng [13] proposed that $\hat{S}_0(t|U = 1)$ decreases from $\hat{S}_0(t_k + 0|U = 1)$ to zero smoothly for all $t > t_k$ and considered the exponential and Weibull distribution functions to complete the tail of the conditional baseline survival function. For the exponential distribution function $\hat{S}_0(t|U = 1) = \exp(-\hat{\zeta}t)$ for $t > t_k$, where $\hat{\zeta}$ satisfies $\exp(-\hat{\zeta}t) = \hat{S}_0(t_k + 0|U = 1)$. For the Weibull distribution function, $\hat{S}_0(t|U = 1) = \exp(-(\hat{\zeta}t)^{\hat{\rho}})$ for $t > t_k$, where $\hat{\zeta}$ and $\hat{\rho}$ are the maximum likelihood estimates based on all observations.

2.3.3. *Variance estimation in semiparametric models*

The standard errors of estimated parameters are not directly available, because of the EM algorithm. The variance of $\hat{\boldsymbol{\gamma}}$ is particularly difficult to compute, because it involves $S_0(t|U = 1)$ trough u_i. Multiple imputation and bootstrap methods have been proposed to estimate the observed information matrix of $\hat{\boldsymbol{\gamma}}$ [6,14]. However, simulations studies (not shown here) indicated that the variance estimated by inverting the Fisher's information matrix of second derivatives when convergence in parameter estimates and likelihood is attained may perform quite well. Non-parametric bootstrap methods [15,16] are implemented in the proposed macro. Resampling from the original dataset is performed through PROC MULLTEST with the BOOTSTRAP option [17]. Computation of bootstrap confidence intervals follows the same conventions of %JACKBOOT macro (http://support.sas.com/ctx/samples/index.jsp?sid=479).

3. Program description

3.1. *Required parameters*

The SAS macro, called PSPMCM, has the basic function of fitting parametric or semiparametric mixture cure models for individual data. The parameters and their description are summurized in Table 1.

The dataset (DATA) is assumed to be entered with one record per individual, with the failure/censoring time (TIME), the censoring indicator (CENSCOD), and covariate vector (VAR). For categorical variables and second or higher order interactions, dummy variables have to be created in a previous data step. In the VAR statement, the name of covariates are separated by blanks. Each variable is followed, into brackets, by the option I, S or IS, which indicates whether it is included as a covariate in the incidence part only (I), the survival part only (S), or both of them (IS). In addition, when plots of survival functions are requested (see below), the values at which the

Table 1 – Parameters required by the SAS macro PSPMCM

Parameter	Description
DATA	SAS data set name to fit the mixture cure model
ID	Subject identification
CENSCOD	Censoring status (1 = event, 0 = censored)
TIME	Failure/censoring time
VAR	E.g.: GENDER(I S,0) AGE(S,10). List of covariates with indicator for incidence (I) or survival (S) and value for survival functions plots
LINK	Link function for incidence regression model
MODEL	Conditional survival function specification
ALPHA	Significance level for confidence interval
BASELINE	Y for output of the conditional baseline survival function
SPLOT	Y for plots of the conditional and marginal survival functions
PLOTFIT	Y for goodness of fit measures
SU0MET	Breslow's or PL method to compute of $\hat{S}_0(t\|U=1)$
TAIL	Tail completion method (ZERO, ETAIL, WTAIL, NONE)
MAXITER	Maximum number of iterations for the EM algorithm
CONVCRIT	Convergence criterium
FAST	Y for output at convergence, N otherwise
BOOTSTRAP	Y for bootstrap methods computation, N otherwise
NSAMPLE	Number of replicates for bootstrap computation
STRATA	Name of the variable for stratified bootstrap resampling
BOOTMET	Type of bootstrap confidence intervals
GESTIMATE	Y for bar charts and Q–Q plots of parameter estimates

survival estimates are plotted are specified after a comma, as shown in the example given in Table 1. Other required statements to specify the model are:

INCPART: to model the incidence part of the model, the user can choose between binary regression models with the logit (LOGIT), probit (PROBIT) or complementary log log (CLOGLOG) links. By default the logistic regression model (LOGIT) is assumed.

SURVPART: specifies the form of the conditional baseline survival function. Parametric models including exponential (EXP), weibull (WEIB), loglogistic (LLOG) and lognormal (LOGN), and the semiparametric Cox's proportional hazards model (COX) are available.

ALPHA: sets the significance level used for the confidence limits for the hazard ratios and odd's ratios. The value must be between 0 and 1. The default value is 0.05, which results in the calculation of a 95% confidence interval.

BASELINE: when set to Y, indicates that the conditional baseline survival function estimate $\hat{S}_0(t|U=1)$, and parameter estimates $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$ and $\hat{\mu}, \hat{\sigma}$ (for parametric models) are written to the BASELINE dataset. If bootstrap resampling is requested for the Cox mixture cure model (see below), the BASELINE_T dataset will moreover contain the estimates for all bootstrap replicates. The default value is N (no output).

SPLOT: when set to Y, the estimated conditional survival curve $\hat{S}_0(t|U=1, \mathbf{x}_i, \mathbf{z}_i)$ and marginal survival curve $\hat{S}_0(t|\mathbf{x}_i, \mathbf{z}_i)$ for individuals with covariate vectors $\mathbf{x}_i$, $\mathbf{z}_i$ are plotted. The values of $\mathbf{x}_i$ and $\mathbf{z}_i$ at which the survival function estimates are computed are specified for each covariate after a comma in the VAR statement. The default value is N.

PLOTFIT: when set to Y, the macro computes for each stratum defined by the covariate vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ the observed (empirical) marginal survival curve $S^{(obs)}(t|\mathbf{x}_i, \mathbf{z}_i)$ which is the Kaplan–Meier estimate for the whole stratum (including right censored subjects). The plots of $(S^{(obs)}(t|\mathbf{x}_i, \mathbf{z}_i), \hat{S}(t|\mathbf{x}_i, \mathbf{z}_i))$ versus t is a visual tool to examine the goodness of the model prediction. The correlation coefficient between $S^{(obs)}(t|\mathbf{x}_i, \mathbf{z}_i)$ and $\hat{S}(t|\mathbf{x}_i, \mathbf{z}_i)$ is also computed for each stratum and provides an appropriate measure of the goodness of fit [18]. *P–P* plots of the KM estimates versus the fitted values are also plotted. The default value is N.

When the Cox PH mixture cure model is requested, additional options are available:

SU0MET: indicates whether the Breslow-type method (CH) or the product limit estimator (PL) is utilized to estimate the conditional baseline survival function. The default value is PL.

TAIL: indicates whether a constraint or a tail completion method is used to estimate $\hat{S}_0(t|U=1)$. The option TAIL = ZERO specifies that the zero tail constraint is utilized (i.e. $\hat{S}_0(t|U=1) = 0$ for $t > t_k$). ETAIL or WTAIL specify that the exponential and Weibull tail completion methods are used, respectively. NONE indicates that no tail constraint is used, but identifiability and convergence problems may rise with this option. The default value is ZERO.

MAXITER: is the maximum number of iterations to perform. If convergence is not attained the displayed output and all output data sets created by the procedure contain results that are based on the last maximum likelihood iteration. The default value is MAXITER = 200.

CONVCRIT: sets the convergence criterion. The default value is 10^{-5}. The iterations are considered to have converged when the maximum relative change in the parameters

Table 2 – Simulation results on biases and MSE of regression parameters

Parameter	Method	True values of $(\beta_0, \beta_1, \gamma_1)$			
		(0,0,0)		(1.3863,−1, −1)	
		Biases	MSE	Biases	MSE
$\hat{\beta}_0$	Weibull mixture	0.0034	0.0194	0.0175	0.0345
	Cox's PH mixture	−0.0019	0.0193	0.0157	0.0337
$\hat{\beta}_1$	Weibull mixture	−0.0216	0.0419	0.0296	0.0628
	Cox' PH mixture	−0.0227	0.0412	0.0301	0.0610
$\hat{\gamma}_1$	Weibull mixture	−0.0070	0.0057	−0.0075	0.0211
	Cox's PH mixture	−0.0107	0.0065	−0.0160	0.0226

and likelihood estimates between iteration steps is less than the value specified.

FAST: when set to Y, parameter estimates and their standard errors (computed by inverting the matrix of second derivatives when convergence is attained) are written to the FAST_INC and FAST_SURV datasets, respectively. Although the standard errors may be underestimated [19], they may be of concern, since they do not require extensive bootstrap computation. The default value is Y.

The following options are available when bootstrap confidence intervals are requested for the Cox's mixture cure model:

BOOTSTRAP: when set to Y, indicates that non-parametric resampling with replacement from the original data set is performed. The default value is N.

NSAMPLE: is the number of bootstrap replicates that are produced by PROC MULLTEST.

STRATA: identifies a single variable to use as a stratification variable in PROC MULLTEST. Stratified resampling may better mimic the observed data.

BOOTMET: specifies the type of bootstrap confidence intervals to compute. These include the percentile (PTCL), hybrid method (HYB), normalized bias corrected (BOOTN), bias corrected (BC) and accelerated bias corrected (BCA) confidence intervals. Jackknife after bootstrap (JACK) is also available. The option BOOTMET = ALL indicates that all methods are requested.

GESTIMATE: when set to Y, Q–Q plots and bar charts of the distribution of parameter estimates over the bootstrap replicates are produced. This allows the user to check graphically the asymptotic normal distribution of parameter estimates and thus to check the validity of the different bootstrap confidence intervals. If the distribution is overdispersed relative to the normal distribution, the validity of the percentile based confidence interval is questionable.

4. Simulation study

In this study, data are generated from a logistic–Weibull mixture cure model, where $\pi(z) = \exp(\beta_0)/[1 + \exp(\beta_0 + \beta_1 z)]$, $S(t|U = 1, z) = \exp[-(\lambda t)^{\rho} \exp(\gamma_1 z)]$. The covariate z is fixed by design and is binary. We set $\lambda = 0.5$ and $\rho = 1$. Censoring times are generated according to the uniform distribution $U[0, 15]$. The results given below are based on $n = 200$ with 500 replications.

Table 2 presents the estimated biases and MSE from the logistic–Weibull and logistic–Cox PH mixture cure models of three regression parameters β_0, β_1 and γ_1 based on simulated data for two different configurations. The first one corresponds to $\pi(z) = 0.5$ that is 50% of the population is cured. In this setting, the covariate z has no effect neither on incidence ($\beta_1 = 0$) nor on latency ($\gamma_1 = 0$). The second configuration corresponds to $\pi(z = 0) = 0.8$ and $\pi(z = 1) = 0.6$, meaning that 20% of the population is cured in one group and 40% in the other. It can be seen that point estimates have little bias. We also simulated a case with a continuous covariate, and similarly there are not substantial departures from the simulated values.

5. Application

The macro is applied to the melanoma data from the Eastern Cooperative Oncology Group (ECOG) phase III clinical trial e1684 [20]. The dataset can be downloaded at http://merlot.stat.uconn.edu/~mhchen/survbook. Our purpose here is not to perform a detailed analysis, but rather to illustrate the use of the PPSMCM macro. An extensive analysis of this dataset is provided in [21,22]. Briefly, the aim of e1684 clinical trial was to evaluate high dose interferon alpha-2b (IFN) regimen against placebo as postoperative adjuvant therapy. After deleting missing informations a total of $n = 284$ are used in the analysis. Only three covariates are included here, both in the incidence and latency parts. These covariates are treatment (0, control group; 1 IFN group), gender (0 for male, 1 for female) and age, which is continuous and centered to the mean. The response variable is taken to be the relapse-free survival (RFS) in years. The Weibull and Cox PH mixture cure models are applied to the data. The cured fraction is modelled by a logistic regression model. For the Cox PH mixture model the zero tail constraint option is specified to compute the conditional baseline survival function estimate. The bootstrap resampling is stratified on the treatment variable, which is the variable of interest. The number of bootstrap replicates is set to 3000 and all the bootstrap confidence intervals computation methods are requested as well as the plot of the distribution of parameter estimates. The measures of goodness of fit are also requested.

For the Cox mixture cure model, the macro is invoked by the following statement:

```
%pspmcm( DATA= e1684, CENSCOD= fcensor, TIME= RFS,
VAR= treatment(I S,1) gender(I S,.)  age (I S,.),
INCPART=logit, SURVPART= Cox, TAIL= zero , SUOMET= pl,
MAXITER= 200, CONVCRIT= 1e-5, ALPHA= 0.05, FAST= Y,
BOOTSTRAP= Y, NSAMPLE= 3000, STRATA= treatment,
BOOTMET= ALL, GESTIMATE= Y,
PLOTFIT= Y)
run;
```

The Kaplan–Meier (KM) plot of the survival function estimate for each treatment group is shown in Fig. 1, together with the estimates from the Weibull and Cox PH mixture cure models. The KM survival function estimate levels off at the right tail and exhibits a long and stable plateau, which ensures the applicability of the mixture cure model approach. The three plots are nearly identical, which allows to think that the model adequacy is good. The correlation coefficients between the KM estimates and the fitted values indicate a good fit for both treatment groups for the Cox mixture cure model ($r = 0.9971$ for both groups). From the Weibull mixture cure model, the fit is better for the treatment group ($r = 0.9984$) than for the control group ($r = 0.9860$).

The parameter estimates from the Weibull and Cox PH mixture cure model are quite similar (Table 3). Moreover, the standard errors from the Cox PH mixture model estimated by using the inverse of the Hessian matrix at convergence of parameter estimates and likelihood are quite similar to those obtained from the Weibull mixture model.

The results indicate that the treatment effect is significant in the incidence (logistic) part, despite a low significance level ($p = 0.040$ for the Weibull model and $p = 0.034$ for the Cox PH

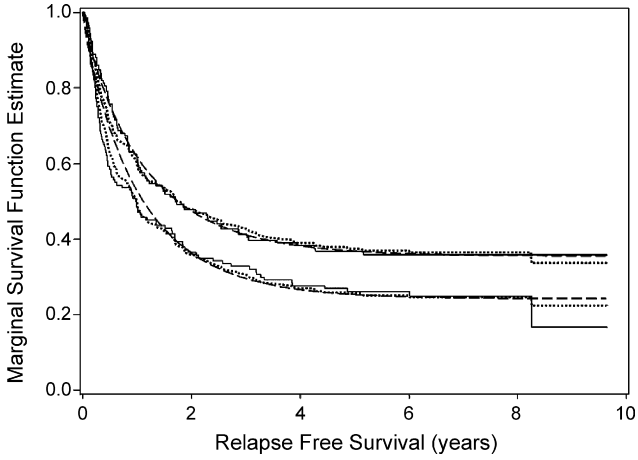

Fig. 1 – Marginal survival function curves for ECOG e1684 data. Upper lines, treatment group; lower lines, control group; solid lines, Kaplan–Meier estimates; dashed lines, Weibull mixture cure model estimates; dotted lines, Cox mixture cure model estimates.

Table 3 – Estimates (standard errors) for the ECOG e1684 data set from the Weibull and Cox's PH mixture cure models

	Weibull mixture model	Cox PH mixture model	
	Estimate (S.E.)[a]	Estimate (S.E.)[b]	BCA 95% CI[c]
Logistic model			
Intercept	1.197(0.239)	1.298(0.236)	0.788; 2.312
Treatment: control	−0.565(0.272)*	−0.574(0.271)*	−1.328; 0.057
Gender: male	−0.061(0.275)	−0.082(0.274)	−0.812; 0.613
Age (per year)	0.014(0.010)	0.018(0.010)	−0.007; 0.062
Survival model			
Treatment: control	−0.104(0.159)	−0.149(0.144)	−0.500; 0.192
Gender: male	0.131(0.152)	0.106(0.147)	−0.268; 0.471
Age (per year)	−0.007(0.006)	−0.007(0.005)	−0.021; 0.006
Intercept (Weibull)	−0.285(0.316)	–	–
Shape (Weibull)	1.088(0.063)	–	–

[a] Standard errors obtained using the inverse of the Hessian matrix.
[b] Standard errors obtained using the inverse of the Hessian matrix at convergence of parameters for the estimates and likelihood.
[c] Bias corrected accelerated bootstrap confidence interval from 3000 replicates.
* p-Value < 0.05.

Table 4 – Estimates (standard errors) for the ECOG e1684 data set from the standard Weibull and Cox PH models

	Standard Weibull model	Standard Cox PH model
	Estimate (S.E.)*	Estimate (S.E.)*
Treatment: control	−0.375(0.144)*	−0.360(0.144)*
Gender: male	0.001(0.148)	−0.018(0.147)
Age (per year)	−0.007(0.005)	0.005(0.005)
Intercept (Weibull)	−0.614(0.115)	–
Shape (Weibull)	0.586(0.035)	–

* p-Value < 0.05.

model), but not in the latency part. This means that the treatment increases the cured fraction, but do not delay the relapse in uncured patients. No other covariate has significant effect, neither on incidence, nor latency.

Invoking the model with the treatment as the only covariate leads to cured fraction estimations of 22.4% (95% confidence interval: 19.1–26.1) for the control group and 33.5% (95% CI: 23.7–44.7) for the IFN group from the Cox PH mixture model. The estimated cured fractions from the Weibull mixture model are 24.2% (95% CI: 20.7–28.1) for the control group and 35.4% (95% CI: 25.5–46.8) for the IFN group.

Applying the standard Weibull and Cox PH survival models (Table 4), would lead to conclude that the treatment had a significant effect on the survival. These models, however, do not account for the possibility of cure and may lead to misinterpretation of covariate effects.

Fig. 2 shows the Q–Q plots for the treatment covariate from the logistic–Cox PH mixture cure model over the 3000 bootstrap replicates. The distribution from the logistic part is overdispersed relative to the normal distribution. Thus, the bias corrected and accelerated bootstrap confidence intervals for parameter estimates is preferred and reported in Table 3. These confidence intervals are slightly larger than those obtained using the variance of parameter estimates. The effect of treatment on incidence is no longer significant (approximated p-value: 0.078).

6. Discussion

We propose a simple SAS macro to estimate parametric and semiparametric mixture cure models for individual data. The proposed SAS macro makes use of standard regression procedures available in the SAS/STAT package since SAS version 7 (PROC NLMIXED is not available in older versions). A limitation of our program is that time dependant variables are not handled, because in this case PROC PHREG does not provide the baseline survival function.

Parametric mixture cure models are simpler to implement than the semiparametric Cox mixture cure model since they do not require the EM algorithm. However, fully-parametric mixture models may not be flexible enough when analyzing biological data, since they involve strong assumptions about the survival distribution function of uncured subjects. A generalized F mixture model has been proposed [27] which makes less distributional assumptions, but computational difficulties may arise.

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 risk that are greater than zero [23,24]. The proportion of individuals considered to

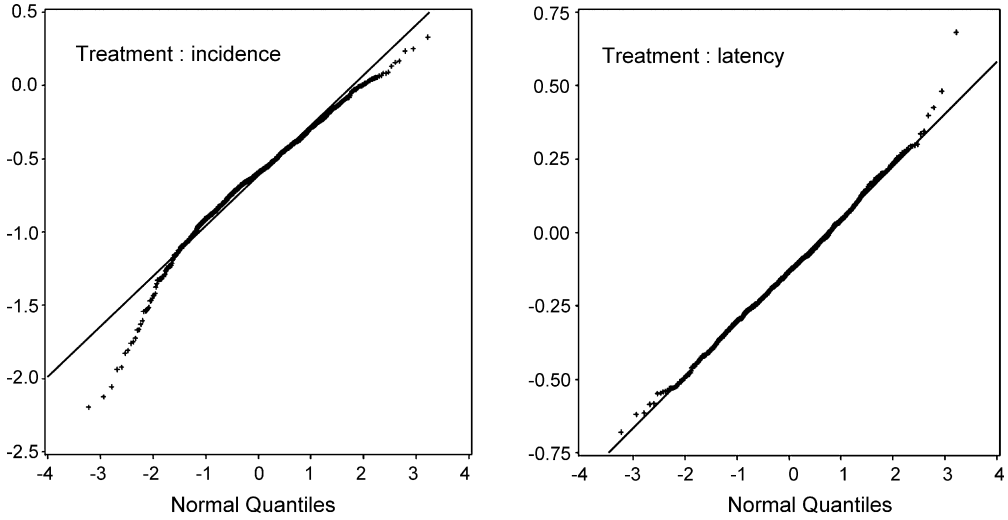

Fig. 2 – Q–Q plots of parameters estimates for the treatment covariate from the Cox mixture cure model based on 3000 bootstrap replicates. Logistic part (left) and survival part (right).

be cured in the former models are generally considered as having a low risk of experiencing the event in the latter models. 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 labelled as $U = 1$ if an event is observed. Usually frailties are assumed to follow a distribution as the gamma, inverse Gaussian or positive stable distribution. These frailty distributions do not allow individuals to have zero risks, hence the standard frailty models do not account for a cured proportion. As an extension of the parametric family, Aalen [25] considered a compound Poisson distribution, which allows a positive probability for the risk to be zero. However, these models only account for heterogeneity among individuals in the latency part but not in the incidence one. Parametric mixture cure models with random effects have recently been proposed [26], but the choice of the frailties' distribution and of their variance matrix may be an important issue.

As stressed by many others [10,28] there are potential problems in applying the mixture cure model in cases where it may not be adequately justified. The use of such models should be restricted to problems in which there are strong biological evidences of the presence of a cured fraction. Another element of caution is that of sufficient follow-up. The levelling off of the Kaplan–Meier curve of the marginal survival function to non-zero proportions, the presence of a long and stable plateau together with a heavy censoring at the tail may also provide graphical evidences of the presence of non-susceptible subjects to ensure the applicability of the mixture cure models [29].

7. Availability

The SAS macro PSPMCM is available to the public at no charge at http://www.isped.u-bordeaux2.fr/recherche/biostats/FR-biostats-accueil.htm# programmes. A simulated dataset is also provided to illustrate the program.

REFERENCES

[1] V.T. Farewell, The use of mixture models for the analysis of survival data with long-term survivors, Biometrics 38 (1982) 1041–1046.

[2] Y. Peng, K.C. Carriere, An empirical comparison of parametric and semiparametric cure models, Biomed. J. 8 (2002) 1002–1014.

[3] J.W. Gamel, E.A. Weller, M.N. Wesley, E.J. Feuer, Parametric cure models of relative and cause-specific survival for grouped survival times, Comput. Methods Programs Biomed. 61 (2000) 99–110.

[4] D.R. Cox, D. Oakes, Analysis of Survival Data, Chapman and Hall, London, New York, 1984, pp. 62–111.

[5] D.R. Cox, Regression models and life-tables (with discussion), J. R. Stat. Soc. Ser. B 34 (1972) 187–220.

[6] Y. Peng, K.B. Dear, A nonparametric mixture model for cure rate estimation, Biometrics 56 (2000) 237–243.

[7] J.P. Sy, J.M. Taylor, Estimation in a Cox proportional hazards cure model, Biometrics 56 (2000) 227–236.

[8] SAS Institute Inc., The NLMIXED Procedure, SAS/STAT User's guide, Version 8, Carry, NC, 2000, pp. 2419–2504.

[9] N.E. Breslow, Contribution to the discussion of D.R. Cox, J. R. Stat. Soc. Ser. B 34 (1972) 216–217.

[10] J.D. Kalbfleisch, R.L. Prentice, The Statiscal Analysis of Failure Time Data, John Wiley, New York, 1980, pp. 84–85.

[11] J.M. Taylor, Semi-parametric estimation in failure time mixture models, Biometrics 51 (1995) 899–907.

[12] C.S. Li, J.M.G. Taylor, J.P. Sy, Identifiability of cure models, Stat. Probab. Lett. 54 (2001) 389–395.

[13] Y. Peng, Estimating baseline distribution in proportional hazards cure models, Comput. Stat. Data Anal. 42 (2003) 187–201.

[14] K.F. Lam, D.Y. Fong, O.Y. Tang, Estimating the proportion of cured patients in a censored sample, Stat. Med. 24 (2005) 1865–1879.

[15] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Application, Cambridge University Press, Cambridge, 1997.

[16] J. Carpenter, J. Bithell, Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, Stat. Med. 19 (2000) 1141–1164.

[17] S.R. Cole, Simple bootstrap statistical inference using the SAS system, Comput. Methods Programs Biomed. 60 (1999) 79–82.

[18] R.A. Maller, S. Zhou, Survival Analysis with Long-term Survivors, Wiley, Chichester, UK, 1996.

[19] H.B. Fang, G. Li, J. Sun, Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model, Scand. J. Stat. 32 (2005) 59–75.

[20] J.M. Kirkwood, M.H. Strawderman, M.S. Ernstoff, T.J. Smith, E.C. Borden, R.H. Blum, Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the Eastern Cooperative Oncology Group Trial EST 1684, J. Clin. Oncol. 14 (1996) 7–17.

[21] J.G. Ibrahim, M.H. Chen, D. Sinha, Bayesian semiparametric models for survival data with a cure fraction, Biometrics 57 (2001) 383–388.

[22] J.G. Ibrahim, M.H. Chen, D. Sinha, Cure Rate Models, Bayesian Survival Analysis, Springer, New York, 2001 (Chapter 5).

[23] O.O. Aalen, Heterogeneity in survival analysis, Stat. Med. 7 (1988) 1121–1137.

[24] P. Hougaard, P. Myglegard, K. Borch-Johnsen, Heterogeneity models of disease susceptibility, with application to diabetic nephropathy, Biometrics 50 (1994) 1178–1188.

[25] O.O. Aalen, Modelling heterogeneity in survival analysis by the compound Poisson distribution, Ann. Appl. Probab. 2 (1988) 951–972.

[26] K.K.W. Yau, A.S.K. Ng, Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma, Stat. Med. 20 (2001) 1591–1607.

[27] Y. Peng, K.B. Dear, J.W. Denham, A generalized f mixture model for cure rate estimation, Stat. Med. 17 (1998) 813–830.

[28] V.T. Farewell, D.A. Sprott, The use of a mixture model in the analysis of count data, Biometrics 44 (1988) 1191–1994.

[29] B. Yu, R.C. Tiwari, K.A. Cronin, E.J. Feuer, Cure fraction estimation from the mixture cure models for grouped survival data, Stat. Med. 23 (2004) 1733–1747.