Approximate Bayesian inference for mixture cure models

E. Lázaro1 · C. Armero1 · V. Gómez-Rubio2

Received: 15 March 2019 / Accepted: 12 September 2019 / Published online: 25 September 2019
© Sociedad de Estadística e Investigación Operativa 2019

Abstract
Cure models in survival analysis deal with populations in which a part of the individuals cannot experience the event of interest. Mixture cure models consider the target population as a mixture of susceptible and non-susceptible individuals. The statistical analysis of these models focuses on examining the probability of cure (incidence model) and inferring on the time to event in the susceptible subpopulation (latency model). Bayesian inference for mixture cure models has typically relied upon Markov chain Monte Carlo (MCMC) methods. The integrated nested Laplace approximation (INLA) is a recent and attractive approach for doing Bayesian inference but in its natural definition cannot fit mixture models. This paper focuses on the implementation of a feasible INLA extension for fitting standard mixture cure models. Our proposal is based on an iterative algorithm which combines the use of INLA for estimating the process of interest in each of the subpopulations in the study, and Gibbs sampling for computing the posterior distribution of the cure latent indicator variable which classifies individuals to the susceptible or non-susceptible subpopulations. We illustrated our approach by means of the analysis of two paradigmatic datasets in the framework of clinical trials. Outputs provide closing estimates and a substantial reduction of computational time in relation to those using MCMC.

Keywords Accelerated failure time mixture cure models · Complete and marginal likelihood function · Gibbs sampling · Proportional hazards mixture cure models · Survival analysis

Mathematics Subject Classification 62F15 · 62N99 · 62N02 · 62P10

1 Department of Statistics and Operations Research, Universitat de València, Valencia, Spain
2 Department of Mathematics, Universidad de Castilla-La Mancha, Albacete, Spain
1 Introduction

Survival analysis is an area of statistics dedicated to researching time-to-event data. This is one of the oldest areas of statistics, which dates back to the 1600s with the construction of life tables. The study of time-to-event data seems simple and traditional because its main interest is focused on nonnegative random variables. But this is very far from being the case. The fact that survival times are always positive keeps it away from the normal distribution framework, censoring and truncation schemes produce non-traditional likelihood issues, and the special elements that generate the dynamic nature of events occurring in time make survival analysis an interesting and exciting area of research and application, mainly in the biomedical field.

Cure models in survival analysis deal with target populations in which a part of the individuals cannot experience the event of interest. This type of models has largely been developed as a consequence of the discovery and development of new treatments against cancer. The rationale of considering a cure subpopulation comes from the idea that a successful treatment removes totally the original tumor and the individual cannot experience any recurrence of the disease. These models allow to estimate the probability of cure, a key and valuable outcome in cancer research. This is not the case for the traditional survival models which consider that all the individuals in the population are at risk. As stated by Lambert et al. (2007), it is important to bear in mind that cure is considered from a statistical, population point of view and not from an individual perspective.

Mixture cure models are the most popular cure models. They consider that the target population is a mixture of susceptible and non-susceptible individuals. The main interest focuses on the so-called incidence model that accounts for the probability of cure, and the latency model for the time to event in the susceptible subpopulation. A mixture model such as this is very attractive, easy to interpret, and allows to account for model complexity (frailties, time-dependent covariates, etc.) in both incidence and latency terms (Peng and Taylor 2014). Some studies in cancer research with this type of models are Sposto (2002) who discussed data from trials in pediatric cancer conducted by the Children’s Cancer Group, Rondeau et al. (2013) who studied recurrences for breast cancer and readmissions for colorectal cancer, and Hurtado Rúa and Dey (2016) who centered on melanoma cancer. A very interesting review of these models up to date is Peng and Taylor (2014). Cured models also appear in other areas of research. This is the case of split population models in economics (Schmidt and Witte 1989) and limited-failure population life models in reliability (Meeker 1987).

Bayesian inference always expresses uncertainty in terms of probability distributions (Loredo 1989, 1992) and uses Bayes’ theorem to update all relevant information. Bayesian methodology is especially attractive for survival analysis due to its natural treatment of censoring and truncation schemes. In addition, Markov chain Monte Carlo (MCMC) methods allow the statistical treatment of complex survival models and the probabilistic quantification of relevant survival outcomes without resorting to asymptotic tools (Ibrahim et al. 2001).

Computational Bayesian statistics focuses on the practical implementation of non-analytical posterior distributions. The integrated nested Laplace approximation (INLA) (Rue et al. 2009) is a recent methodology for doing approximate Bayesian
inference in the framework of latent Gaussian models (LGM) (Rue and Held 2005). These models are a special class of Bayesian additive models that cover a wide range of studies and applications (Rue et al. 2017), and survival models in particular (Martino et al. 2011). INLA, in comparison with MCMC methods, provides accurate and fast approximations to the relevant posterior marginal distributions.

INLA is very attractive and has very good properties, but it also has some limitations. In particular, INLA cannot fit mixture models (Marin et al. 2005) in a natural way because they are generally defined in terms of a combination of different distributions (Gómez-Rubio and Rue 2018). But in science, every constraint or difficulty becomes an opportunity for learning. On this matter, Bivand et al. (2014) and Gómez-Rubio and Rue (2018) propose the combination of INLA within MCMC for mixture models: the Metropolis–Hastings algorithm and INLA for the relevant posterior conditional distributions. Gómez-Rubio (2018) extends these proposals and introduce Modal Gibbs sampling to accelerate the inferential process.

This paper focuses on the implementation of INLA in mixture cure models. A general mixture cure survival model with covariate information for the latency and the incidence model within a general scenario with censored and non-censored information is discussed. The fact that non-censored individuals undoubtedly belong to the uncured population is a valuable information that is incorporated in the general inferential process. Recently, INLA has implemented a mixture cure model that does not allow for the inclusion of covariates in the cure fraction and only considers the Weibull distribution to describe the latency model [Documented in the INLA package and accessible through inla.doc("cure"), Rue et al. (2009)].

The organization of this paper is as follows. Section 2 presents the main elements of mixture cure models and the two most popular mixture cure models, the Cox proportional hazards and the accelerated failure times models. Section 3 introduces the integrated nested Laplace approximation within the general framework of Bayesian inference. Section 4 is the core of the paper and contains our INLA proposal for estimating mixture cure models. Section 5 applies our proposal to the statistical analysis of two benchmark data sets in the framework of clinical trials and bone marrow transplants, and discusses and compares the subsequent results with those from a MCMC implementation. The paper ends with some conclusions.

2 Mixture cure models

Let $T^*$ be a continuous and nonnegative random variable that describes the time to event of an individual in some target population. Let $Z$ be a cure random variable defined as $Z = 0$ if that individual is susceptible for experiencing the event of interest, and $Z = 1$ if she/he is cured or immune for that event. Cure and non-cure probabilities are $P(Z = 1) = \eta$ and $P(Z = 0) = 1 - \eta$, respectively. The survival function for individuals in the cured and uncured population, $S_c(t)$ and $S_u(t)$, $t > 0$, respectively, is

$$S_u(t) = P(T^* > t \mid Z = 0)$$
The general survival function for $T^*$ can be expressed in terms of a mixture of both cured and uncured populations in the form

$$S(t) = \Pr(T^* > t) = \eta + (1 - \eta) S_u(t).$$

(2)

It is important to point out that $S_u(t)$ is a proper survival function but $S(t)$ is not. It goes to $\eta$ and not to zero when $t$ goes to infinity. Cure fraction $\eta$ is also known as the incidence model and time to event $T_u^*$ in the uncured population as the latency model (Peng and Taylor 2014).

**Covariates in the incidence model**

The effect of a baseline covariate vector $x_1$ on the cure proportion is typically modeled by means of a logistic link function, logit[$\eta(\beta_1)$] = $\beta_1'x_1$, also expressed as

$$\eta(\beta_1) = \frac{\exp(\beta_1'x_1)}{1 + \exp(\beta_1'x_1)},$$

(3)

where $\beta_1$ is the vector of regression coefficients associated to $x_1$. Note that other link functions can be used to connect the cure fraction with the vector of covariates $x_1$ such as the probit link or the complementary log–log link (see Robinson (2014) for more details).

**Covariates in the latency model**

The most common regression models in survival analysis are the Cox proportional hazards model (Cox 1972) and the accelerated failure time models. We will introduce them below.

**Cox proportional hazards model, CPH.** It is usually formulated in terms of the hazard function for the time to event $T_u^*$, or instantaneous rate of occurrence of the event, as

$$h_u(t \mid h_{u0}, \beta_2) = \lim_{\Delta t \to \infty} \frac{\Pr(t \leq T_u^* < t + \Delta t \mid T_u^* \geq t)}{\Delta t} = h_{u0}(t) \exp(\beta_2'x_2),$$

(4)

where $h_{u0}(t)$ is the baseline hazard function that determines the shape of the hazard function. Model (4) can also be presented in terms of the survival function of $T_u^*$ as

$$S_u(t \mid S_{u0}, \beta_2) = [S_{u0}(t)]^{exp[\beta_2'x_2]},$$

(5)

where $S_{u0}(t) = \exp{- \int_0^t h_{u0}(s)ds}$ represents the survival baseline function.
Fully Bayesian methods specify a model for \( h_{u0}(t) \) which may be of parametric or nonparametric nature. Exponential, Weibull, and Gompertz hazard functions are common parametric proposals in the empirical literature. Mixture of piecewise constant functions or B-splines basis functions are the usual counterpart in nonparametric selections. They provide a great flexibility to the modeling by allowing different patterns and multimodalities, but some care is needed when working with them to avoid overfitting. To this effect, the elicitation of prior distributions is a relevant issue in the Bayesian approach to regularization (Lázaro et al. 2018).

Accelerated failure time models, AFT. These models try to adapt the philosophy of linear models to the survival framework. The survival variable \( T_u^* \) is now expressed in the logarithmic scale to extend the modeling to the real line. It is modeled as the sum of a linear term for the covariates \( x_2 \), which usually includes an intercept element, and a random error \( \epsilon \) amplified or reduced by a scale factor \( \sigma \) as follows:

\[
\log(T_u^*) = x'_2 \beta_2 + \sigma \epsilon. \tag{6}
\]

Common distributions for \( \epsilon \) are normal, logistic, and standard Gumbel. They, respectively, imply log-normal, log-logistic and Weibull distributions for \( T_u^* \) (Christensen et al. 2011). Weibull AFT models are the most popular ones, in which covariates \( x_2 \) are commonly included in the scale parameter as

\[
\lambda(\beta_2) = \exp\{\beta'_2 x_2\},
\]

and consequently

\[
h_u(t \mid \alpha, \beta_2) = \alpha t^{\alpha-1} \exp[\beta' x_2],
\]

\[
S_u(t \mid \alpha, \beta_2) = \exp\{-t^\alpha \exp[\beta'_2 x_2]\}. \tag{7}
\]

This modeling strategy based on introducing covariate information through one of the parameters of the target distribution also applies to the rest of parametric probability distributions.

### 3 Bayesian inference and the integrated nested Laplace approximation

Bayesian inference derives the posterior distribution of the quantities of interest according to Bayes’ theorem, which combines the prior distribution of all unknown quantities and the likelihood function constructed from the data. It is the main element in Bayesian statistics and starting point of all relevant inferences. The posterior distribution in complex models is non-analytical, and for this reason, it needs to be computationally approached. To that effect, MCMC methods are surely the most popular procedures although they involve large computational costs and require additional work for checking convergence and accuracy estimation.

The structure and main elements of the INLA approach for doing Bayesian inference are summarized below. Let us assume a set of \( n \) random variables \( T^* = (T^*_1, \ldots, T^*_n) \) mutually conditionally independent given a latent Gaussian Markov random field (GMRF) (Rue and Held 2005) \( \theta \) and a set of likelihood hyperparameters \( \phi_2 \). The
GMRF $\theta$ depends on some hyperparameters $\phi_1$ and can include effects of different type (regression coefficients, random effects, seasonal effects, etc).

According to Bayes’ theorem, the joint posterior distribution for $(\theta, \phi)$, where $\phi = (\phi_1, \phi_2)$, after data $D = \bigcup_{i=1}^{n} D_i$ have been observed, where $D_i$ represents the data from individual $i$th, can be written as

$$
\pi(\theta, \phi | D) \propto \prod_{i=1}^{n} L_i(\theta, \phi | D) \pi(\theta, \phi)
$$

where $L_i(\theta, \phi | D)$ is the likelihood function of $(\theta, \phi)$ for data $D_i$, and $\pi(\theta, \phi)$ represents the prior distribution of $(\theta, \phi)$ which factorizes as the product of a GMRF conditional prior distribution $\pi(\theta | \phi)$ and a marginal prior distribution $\pi(\phi)$.

INLA makes use of Laplace approximations (Rue et al. 2009) to obtain approximations $\tilde{\pi}(\phi | D)$ and $\tilde{\pi}(\theta | \phi, D)$ for the posterior distribution $\pi(\phi | D)$ and $\pi(\theta | \phi, D)$, respectively, where $\theta$ denotes a generic univariate element in $\theta$. The marginal posterior distribution for the latent terms $\pi(\theta | D)$ can be obtained as

$$
\pi(\theta | D) = \int \pi(\theta | \phi, D) \pi(\phi | D) \, d\phi,
$$

and consequently, it can be approximated by numerical integration as

$$
\tilde{\pi}(\theta | D) \approx \sum_m \tilde{\pi}(\theta | \phi_m, D) \tilde{\pi}(\phi_m | D) \Delta_m,
$$

where $\phi_m$ are points in the hyperparametric space $\Phi$, and $\Delta_m$ integration weights.

The posterior marginal distribution $\pi(\phi | D)$ can also be approximated by numerical integration according to the expression

$$
\pi(\phi | D) = \int \pi(\phi | D) \, d\phi_{\phi - \phi},
$$

where $\phi_{\phi - \phi}$ represents all elements in $\phi$ except $\phi_\phi$.

INLA is implemented in the package R-inla for the R statistical software (Core Team 2014). This package implements a number of latent effects and allows for an easy model fitting and visualization of the output. A recent review on INLA can be found in Rue et al. (2017).

4 INLA to estimate mixture cure models

In general, standard survival models such as CPH and AFT models can be expressed in terms of GMRF models, and consequently, they can be adapted for its INLA implementation (Akerkar et al. 2010; Martino et al. 2011). In the case of CPH models, the baseline hazard function is reparameterized in the exponential scale in order to
be included in the CPH element that accounts for regression information. This exponential term also allows the inclusion of time-varying covariate effects, nonlinear, structured or non-random effects, spatial modeling, etc (Hennerfeind et al. 2006). They can be expressed by means of a structured geoadditive predictor whose elements can be modeled in terms of a GMRF model. AFT models also have this nice relationship and behavior for INLA implementation.

Gibbs sampler for mixture estimation

Let us consider a general survival scenario in the framework of non-informative and independent right censoring and a mixed cure sampling model. Survival time is defined as the pair \((T, \delta)\), where \(T = \min(T^*, C)\), \(C\) being the censoring time, and \(\delta\) an indicator function defined as \(\delta = 0\) when the subsequent observation is censored \((T^* > C)\), and \(\delta = 1\) when it is not. We assume that the distribution of \(T^*\) depends on a conditional GRMF \(\theta\) on hyperparameters \(\phi_1\) and a likelihood hyperparametric vector \(\phi_2\), and consider \(\pi(\theta, \phi)\) as the prior distribution for \((\theta, \phi)\) which factorizes as

\[
\pi(\theta, \phi) = \pi(\theta | \phi) \pi(\phi).
\] (12)

Let \(D_i = (t_i, \delta_i)\) represent the survival observed data for individual \(i, i = 1, \ldots, n\), and \(D = \bigcup_{i=1}^n D_i\). The complete data for individual \(i\) is defined as \(D_{\text{com}, i} = (t_i, \delta_i, z_i) = (D_i, z_i)\), which includes the value \(z_i\) of the subsequent latent variable that classifies this individual as cured or not, and \(D_{\text{com}} = \bigcup_{i=1}^n D_{\text{com}, i}\). It should be noted that an observed survival time clearly indicates that the subsequent individual belongs to the uncured population.

The complete data likelihood function is the product of the complete likelihood function for each individual defined as Ibrahim et al. (2001)

\[
\mathcal{L}(\theta, \phi | D_{\text{com}}) = \prod_{i=1}^n \mathcal{L}_i(\theta, \phi | D_{\text{com}}) = \prod_{i=1}^n \eta_i(\theta, \phi) z_i (1 - \eta_i(\theta, \phi))^{1 - z_i} h_{iu}(t_i | \theta, \phi) \delta_i (1 - z_i) S_{iu}(t_i | \theta, \phi)^{(1 - z_i)}.
\] (13)

As \(z\) is seldom observed, it is often treated as another parameter in the model and its posterior distribution needs to be computed as well. The posterior distribution for \((\theta, \phi)\) computed from Bayes’ theorem would be

\[
\pi(\theta, \phi | D) \propto \mathcal{L}(\theta, \phi | D) \pi(\theta, \phi) \propto \sum_{z \in \mathcal{Z}} \mathcal{L}(\theta, \phi | D_{\text{com}}) \pi(\theta, \phi),
\]

where \(\mathcal{L}(\theta, \phi | D)\) is the likelihood function of \((\theta, \phi)\) for the observed data \(D\), and \(\mathcal{Z}\) denotes the parameter space of the cure indicator values, which is the \(n\)-dimensional Cartesian product of the binary set \(\{0, 1\}\).
The introduction of the latent indicator in the inferential process and the Gibbs sampler is the usual procedure to approach Bayesian mixture estimation (Diebolt and Robert 1994; Marin et al. 2005). We follow this proposal and consider the inferential process defined by the joint posterior distribution

\[ \pi(\theta, \phi, z \mid D) \propto \mathcal{L}(\theta, \phi \mid D_{\text{com}}) \pi(\theta, \phi), \]

and a Gibbs sampler based on the full conditional posterior distributions \( \pi(\theta, \phi \mid D_{\text{com}}) \) and \( \pi(z \mid \theta, \phi, D) \).

**INLA and modal Gibbs**

Our proposal for fitting mixture cure models by means of INLA is based on Gómez-Rubio and Rue (2018) and Gómez-Rubio (2018). They use INLA for estimating the conditional posterior marginals of the model parameters \( \pi(\theta \cdot \mid D_{\text{com}}) \) and \( \pi(\phi \cdot \mid D_{\text{com}}) \), which assumes that the latent vectors \( z \) which determine the subpopulation to which each individual belongs to are known. All relevant marginal posterior distributions, \( \pi(\theta \cdot \mid D_{\text{com}}) \) and \( \pi(\phi \cdot \mid D_{\text{com}}) \), can be fitted as usual in the INLA approach. Note that this involves actually fitting two models with INLA given the latent variable \( z \): one with a binomial likelihood and a survival model (as in Martino et al. 2011).

The posterior marginal distribution for each \( \theta \cdot \) can be computed as

\[
\pi(\theta \cdot \mid D) = \sum_{z \in Z} \pi(\theta \cdot, z \mid D) = \sum_{z \in Z} \pi(\theta \cdot \mid D_{\text{com}}) \pi(z \mid D),
\]  

(14)

where \( \pi(\theta \cdot \mid D_{\text{com}}) \) is fitted by INLA and \( \pi(z \mid D) \) is the marginal posterior distribution for the latent cure indicator vector based on the observed data. This latter distribution will be computed using modal Gibbs sampling as proposed by Gómez-Rubio (2018). The computation of \( \pi(\phi \cdot \mid D) \) follows a similar procedure.

Expression (14) needs some additional discussion so that it can be better adapted to the cure models framework. Here, we know that each survival observation can be censored or uncensored. In the case of a censored data, we do not know if the subsequent individual can or cannot experience the event of interest, hence their belonging to the uncured or cured subpopulation is unknown, and consequently, there will be uncertainty about the value of the corresponding cure indicator variable. Conversely, an uncensored observation will indicate that the subsequent individual has surely experienced the event of interest, and therefore, she/he belongs to the uncured subpopulation. If we split \( z = (z_{\text{unc}}, z_{\text{cen}}) \), where \( z_{\text{unc}} \) (\( z_{\text{cen}} \)) represents the \( n_{\text{unc}} \) (\( n_{\text{cen}} \))-dimensional latent cure indicator corresponding to the uncensored (censored) data, the complete knowledge on the value of the latent indicator of the uncensored data will imply
\[ z_{\text{unc}} = 0. \] For this reason,
\[
\pi(z \mid D) = \pi(z_{\text{unc}}, z_{\text{cen}} \mid D) = \begin{cases} 
\pi(z_{\text{cen}} \mid D) & \text{for } z_{\text{unc}} = 0, z_{\text{cen}} \in Z_{\text{cen}} \\
0 & \text{otherwise},
\end{cases}
\]
where now \( Z_{\text{cen}} \) is the parameter space of the cure indicator variables for the censored observations, with lower dimensionality than \( Z \). Hence, expression (14) can be rewritten as
\[
\pi(\theta, D) = \sum_{z_{\text{cen}} \in Z_{\text{cen}}} \pi(\theta, D_{\text{com}}) \pi(z_{\text{cen}} \mid D) . \tag{15}
\]

The above procedure can be described in a more structured way via the following algorithm:

**Step 0.** Assign initial values to the latent cure indicator of the \( n_{\text{cen}} \) censored observations, \( z^{(0)}_{\text{cen}} \), and consider \( z_{\text{unc}} = 0 \) for the uncensored observations. Define \( z^{(0)} = \{z^{(0)}_{\text{cen}}, z_{\text{unc}}\} \).

**Step 1.** For \( m = 1, 2, \ldots \)

(a) Use INLA to approximate \( \pi(\theta, z^{(m-1)}, D) \) and \( \pi(\phi, z^{(m-1)}, D) \).

(b) Compute the subsequent posterior (conditional) modes \( \hat{\theta}^{(m)} \) and \( \hat{\phi}^{(m)} \), respectively, from each of the posterior distributions in (a).

(c) Sample \( z^{(m)}_{\text{cen}} = (z^{(m)}_{\text{cen},1}, \ldots, z^{(m)}_{\text{cen},n_{\text{cen}}}) \) from the full conditional distribution for the cure latent variable (Marin et al. 2005; Cai et al. 2012)
\[
\pi(Z = 0 \mid D, \hat{\theta}^{(m)}, \hat{\phi}^{(m)}) = \frac{(1 - \eta(\hat{\theta}^{(m)}, \hat{\phi}^{(m)})) S_u(t \mid \hat{\theta}^{(m)}, \hat{\phi}^{(m)})}{\eta(\hat{\theta}^{(m)}, \hat{\phi}^{(m)}) + (1 - \eta(\hat{\theta}^{(m)}, \hat{\phi}^{(m)})) S_u(t \mid \hat{\theta}^{(m)}, \hat{\phi}^{(m)})} , \\
\pi(Z = 1 \mid D, \hat{\theta}^{(m)}, \hat{\phi}^{(m)}) = 1 - \pi(Z = 0 \mid D, \hat{\theta}^{(m)}).
\]

(d) Define \( z^{(m)} = \{z^{(m)}_{\text{cen}}, z_{\text{unc}}\} \).

At this point, it is important to comment that the computational complexity in expression (14) comes from the great number of summands involved. In this sense, a practical idea for reducing computational times of the algorithm is to avoid the recalculation of the outputs corresponding to configurations in Step 1.(c) that have already been explored. That is, if a model given a particular \( z \) has already been fit, it is not necessary to fit it again when the same value of the latent variable is drawn.

### 5 Illustrative studies

We considered two benchmark datasets to illustrate our proposal for estimating mixture cure models via INLA. They are the so-called Eastern Cooperative Oncology Group (ECOG) phase III clinical trial e1684 dataset (Kirkwood et al. 1996) and the bone
marrow transplant study dataset (Kersey et al. 1987). In both studies, we compared our results with the ones obtained via MCMC methods. Inferences in both studies were performed on a Windows laptop with an Intel(R) Core(TM) i5-52004 2.20GHz processor. All implementations were made in the R environment (version 3.4.3), and the code is provided on Github at https://github.com/becarioprecario/cure_rate_models_INLA. We used the R-INLA package for INLA and JAGS software (version 4.3.0) through the rjags package (Plummer 2003) for MCMC inferences.

ECOG study

The ECOG phase III clinical trial was designed to compare a high-dose interferon alpha-2b (IFN) regimen against close observation which was the standard therapy (ST) as the postoperative adjuvant treatment (Kirkwood et al. 1996) in high-risk melanoma patients. Data in the analysis included a total of 284 observations, of which 88 were right-censored. Relapse-free survival (FFS), in years, was one of variables of interest in the study and now our survival variable. Covariate information included gender, 113 women (W) and 171 men (M), treatment (144 people in the IFN group and 140 in ST), and age (A) (in years and centered on the sample mean). FFS sample median was 1.24 and 1.36 years in the case of M and W, and 1.82 and 0.98 years in the IFN group and ST, respectively.

Incidence and latency model.

We considered the same CPH mixture cure model stated by the authors in (Kirkwood et al. 1996). The cure proportion for individual i in the incidence model was expressed in terms of a binary regression logistic model defined as

\[
\text{logit}[\eta_i(\beta_1)] = \beta_{0,1} + \beta_{W,1} I_W(i) + \beta_{IFN,1} I_{IFN}(i) + \beta_{A,1} A_i,
\]

where \(\beta_{0,1}\) represents the reference category, to be a man receiving ST treatment, and \(I_G(i)\) is an indicator variable with value 1 if individual i has the characteristic \(G\) and 0 otherwise.

Survival time for individual i the uncured subpopulation was modeled by a CPH model with hazard function,

\[
h_{ui}(t | h_{u0}, \beta_2) = h_{u0}(t) \exp\{\beta_{0,2} + \beta_{W,2} I_W(i) + \beta_{IFN,2} I_{IFN}(i) + \beta_{A,2} A_i\},
\]

with Weibull baseline hazard function \(h_{u0}(t) = \alpha t^{\alpha - 1}\).

The model is completed with the elicitation of a prior distribution for all uncertainties it includes. We assume prior independence and select vague normal distributions centered at zero and variance 1000 for all the regression coefficients in (16) and (17) as well as for \(\log(\lambda)\). The elicited prior distribution for \(\alpha\) is the gamma distribution \(\text{Ga}(0.01, 0.01)\), a very common election in these models which baseline hazard function is specified in terms of a Weibull distribution.
Posterior inferences

Our algorithm configuration included 50 burn-iterations followed by other 450 iterations for inference. In addition, the simulations were thinned by storing one in five draws in order to reduce autocorrelation in the saved sample. The convergence was evaluated by examining whether the estimated conditional (on \( z \)) marginal log-likelihood achieved stability during the iteration steps of the algorithm.

INLA results were compared to those obtained via MCMC methods with the JAGS software. The MCMC algorithm ran for three Markov chains with 100,000 iterations after a burn-in period with 20,000 iterations. In addition, the chains were thinned by storing one in two hundred iterations in order to reduce autocorrelation in the subsequent sample and save storage space. Convergence was assessed based on the potential scale reduction factor, and the effective number of independent simulation draws (Gelman and Rubin 1992). It is remarkable that to perform both analyses the response variable (Relapse-free survival (FFS), in years) was scaled by dividing each observation by the maximum observed value to avoid INLA numerical overflow in latency model computation.

The number of iterations needed to accomplish convergence under our proposal is a fraction than the one in the MCMC configuration. Furthermore, the parameters space is not explored as their posterior marginals are computed from the conditional posterior marginals obtained with INLA. Computational times with INLA get reliable estimates in 28 min and the MCMC sampler needed around 55 min.

Table 1 shows a summary of the INLA and MCMC approximate posterior marginal distribution of the parameters of the mixture cure model estimated. The agreement in all the outputs is quite high and confirms that our approach works and provides similar estimates to MCMC.

The estimation of the cure proportion and the survival profiles for the different groups of individuals are relevant issues in the medical context of the study. INLA computes an approximation to the conditional marginal log-likelihood function \( \pi(D|z) \), and it can be used to select the most likely configuration of the latent vector \( z \) that has been generated during the sample process to approximate the posterior distribution of the cure proportion and the survival profiles. In particular, the inla.posterior.samples function in the R-INLA package may be used to generate samples from the approximated joint posterior distribution of the estimated model (we select the most likely model). Additionally, these samples can subsequently be processed to derive approximated posterior distributions for the quantities of interest.

Table 2 includes the INLA and MCMC posterior mean, standard deviation, and 95% credible interval of the posterior distribution of the cure proportion for individuals in the four groups of interest: men treated with the standard therapy (M-ST), men treated with interferon alpha-2b (M-IFN), women in the standard therapy group (W-ST), and IFN women (W-IFN). Outcomes from INLA and MCMC also are in close agreement and highlight that W-IFN individuals present the highest cure proportion estimates, while the lowest values correspond to M-ST ones. Differences between treatments are clinically relevant for both women and men.
Table 1  Summary of the INLA and MCMC approximate marginal posterior distributions: mean, standard deviation, 95% credible interval, and posterior probability that the subsequent parameter is positive

| Parameter       | Mean  | Sd    | 95 % CI          | P(> 0) |
|-----------------|-------|-------|------------------|--------|
| **Incidence**   |       |       |                  |        |
| INLA            |       |       |                  |        |
| $\beta_{0,1}$   | $-1.203$ | $0.234$ | $[-1.676,-0.759]$ | $0.000$ |
| $\beta_{W,1}$  | $0.064$  | $0.274$ | $[-0.476,0.599]$  | $0.595$ |
| $\beta_{IFN,1}$ | $0.578$  | $0.271$ | $[0.050,1.113]$   | $0.984$ |
| $\beta_{A,1}$  | $-0.015$ | $0.010$ | $[-0.036,0.006]$  | $0.079$ |
| MCMC            |       |       |                  |        |
| $\beta_{0,1}$   | $-1.219$ | $0.241$ | $[-1.711,-0.776]$ | $0.000$ |
| $\beta_{W,1}$  | $0.062$  | $0.282$ | $[-0.474,0.603]$  | $0.580$ |
| $\beta_{IFN,1}$ | $0.567$  | $0.278$ | $[0.003,1.131]$   | $0.975$ |
| $\beta_{A,1}$  | $-0.015$ | $0.011$ | $[-0.036,0.007]$  | $0.086$ |
| **Latency**     |       |       |                  |        |
| INLA            |       |       |                  |        |
| $\alpha$        | $0.918$  | $0.052$ | $[0.818,1.022]$   | $-$    |
| $\exp(\beta_{0,2})$ | $7.591$ | $1.126$ | $[5.581,9.990]$   | $-$    |
| $\beta_{W,2}$  | $0.131$  | $0.158$ | $[-0.184,0.437]$  | $0.796$ |
| $\beta_{IFN,2}$ | $-0.096$ | $0.154$ | $[-0.401,0.206]$  | $0.266$ |
| $\beta_{A,2}$  | $-0.007$ | $0.006$ | $[-0.018,0.004]$  | $0.098$ |
| MCMC            |       |       |                  |        |
| $\alpha$        | $0.908$  | $0.053$ | $[0.907,1.013]$   | $-$    |
| $\exp(\beta_{0,2})$ | $7.282$ | $1.162$ | $[5.223,9.764]$   | $-$    |
| $\beta_{W,2}$  | $0.131$  | $0.164$ | $[-0.193,0.425]$  | $0.786$ |
| $\beta_{IFN,2}$ | $-0.111$ | $0.163$ | $[-0.439,0.204]$  | $0.240$ |
| $\beta_{A,2}$  | $-0.008$ | $0.006$ | $[-0.018,0.003]$  | $0.086$ |

Table 2  Summary of the approximated INLA and MCMC posterior mean, standard deviation, and 95% credible interval of the cure proportion for averaged age individuals in the four groups of the study

|       | Group    | Mean  | Sd    | 95 %CI          |
|-------|----------|-------|-------|------------------|
| **INLA** |          |       |       |                  |
| $M-ST$ | 0.244    | 0.041 | [0.242, 0.330] |
| $M-IFN$| 0.364    | 0.046 | [0.277, 0.455] |
| $W-ST$ | 0.258    | 0.047 | [0.176, 0.359] |
| $W-IFN$| 0.381    | 0.055 | [0.278, 0.493] |
| **MCMC** |         |       |       |                  |
| $M-ST$ | 0.231    | 0.042 | [0.153, 0.315] |
| $M-IFN$| 0.344    | 0.048 | [0.253, 0.442] |
| $W-ST$ | 0.243    | 0.050 | [0.156, 0.346] |
| $W-IFN$| 0.359    | 0.058 | [0.252, 0.476] |

Figure 1 displays the INLA and MCMC mean of the posterior distribution of the uncured survival function for individuals in each of the four groups of interest. Estimation from both approaches scarcely differs. From a clinical point of view, the survival profiles are very similar among the groups, but it seems that the best and worst survival expectations correspond to $M-IFN$ and $W-ST$ groups, respectively. We could conclude that the probability of cure is very different among the groups (see Table 2), but the
The uncured survival profiles of the individuals in the different groups are very similar (see Fig. 1).

**Bone marrow transplant study**

Next, we consider the bone marrow transplant study dataset in Kersey et al. (1987) to illustrate our proposal for a Weibull AFTMC model. This study was undertaken to compare autologous and allogeneic marrow transplantation with regard to survival times of patients affected with lymphoblastic leukemia and poor prognosis. A total of 91 patients were treated with high-doses of chemoradiotherapy and followed-up during a period between 1.4 and 5.0 years. Forty-six patients with a HLA-matched donor received a donor marrow (allogeneic graft) and 45 patients without a matched donor received their own marrow taken during remission and purged of leukemic cells with the use of monoclonal antibodies (autologous graft). The survival variable of interest was time to death, in days, which ranged from 11 to 1845 days. Data contain 22 right-censored observations and 69 uncensored. In general, times to death are longer for allogeneic transplant patients (sample median was 292 days) than for autologous patients (sample median was 112 days).
The main goal of the study was to compare both groups, autologous and allogeneic, with regard to the incidence and the latency models. Covariate information only contemplates the type of transplant and was incorporated in both terms of the cure model.

**Incidence and latency model**

The cure probability for individual $i$th corresponding to the incidence model was expressed in terms of a regression logistic model defined as

$$\text{logit} [\eta_i(\beta)] = \beta_{\text{All},1} + \beta_{\text{Aut},1} I_{\text{Aut}}(i),$$

where $\beta_{\text{All},1}$ represents the effect of the reference category, to be an individual who has received an allogeneic transplant, and $I_{\text{Aut}}(i)$ is an indicator variable with value 1 whether individual $i$ has had an autologous graft.

Survival time for individual $i$ in the uncured subpopulation, $T_{ui}$, was modeled by means of a Weibull AFT model defined as

$$\log(T_{ui}) = \beta_{\text{All},2} + \beta_{\text{Aut},2} I_{\text{Aut}}(i) + \sigma \epsilon_i,$$

where now $\beta_{\text{All},2}$ represents the effect of receiving an allogeneic graft, and $\beta_{\text{Aut},1}$ the additional effect for having an autologous transplant.

The model is completed with the elicitation of a prior distribution for all parameters it contains. We assume prior independence and select vague normal distributions centered at zero and variance 1000 for all the regression coefficients in the model except for $\alpha = 1/\sigma$, for which a Ga(0.1, 0.1) distribution was selected.

**Posterior inferences**

Our algorithm configuration for this model included 20 burn-in iterations and other 180 for inference. In addition, the simulations were thinned by storing every 2nd draws.

**Table 3** Summary of the INLA and MCMC approximate marginal posterior distributions: mean, standard deviation, 95% credible interval, and posterior probability that the subsequent parameter is positive

| Parameter | INLA Mean | Sd | 95% CI | $P(\cdot > 0)$ |
|-----------|-----------|----|--------|---------------|
| Incidence |           |    |        |               |
| $\beta_{\text{All},1}$ | -0.961    | 0.333 | [-1.647, -0.336] | 0.000 |
| $\beta_{\text{Aut},1}$ | -0.433    | 0.501 | [-1.432, 0.535] | 0.192 |
| MCMC      |           |    |        |               |
| $\beta_{\text{All},1}$ | -1.016    | 0.344 | [-1.799, -0.392] | 0.000 |
| $\beta_{\text{Aut},1}$ | -0.413    | 0.522 | [-1.450, 0.670] | 0.204 |
| Latency   |           |    |        |               |
| $\beta_{\text{All},2}$ | 2.222     | 0.236 | [1.748, 2.672] | 1.000 |
| $\beta_{\text{Aut},2}$ | 0.736     | 0.261 | [0.226, 1.248] | 0.998 |
| $\alpha$  | 1.143     | 0.102 | [0.947, 1.347] | -    |
| MCMC      |           |    |        |               |
| $\beta_{\text{All},2}$ | 2.151     | 0.264 | [1.616, 2.618] | 1.000 |
| $\beta_{\text{Aut},2}$ | 0.735     | 0.269 | [0.207, 1.246] | 1.000 |
| $\alpha$  | 1.123     | 0.102 | [0.931, 1.323] | -    |
Table 4  Summary of the approximated INLA and MCMC posterior mean, standard deviation, and 95% credible interval of the cure proportion for allogeneic and autologous graft patients

| Group | Mean | Sd  | 95 % CI       |
|-------|------|-----|---------------|
| INLA  | All  | 0.287 | 0.067         | [0.171,0.428] |
|       | Aut  | 0.204 | 0.060         | [0.105,0.338] |
| MCMC  | All  | 0.271 | 0.065         | [0.142,0.403] |
|       | Aut  | 0.200 | 0.062         | [0.099,0.341] |

Fig. 2  Posterior mean of the uncured survival function for Allogeneic (on the left) and Autologous (on the right) transplanted patients computed from INLA (black solid line) and MCMC (gray dashed line) in order to reduce autocorrelation in the saved sample. Convergence was evaluated by examining whether the conditional marginal log-likelihood estimates achieved stability during the iteration steps of our algorithm.

MCMC simulation was run considering three Markov chains with 200,000 iterations and a burn-in period with 40,000 iterations. The chains were thinned by storing every 400th iteration to reduce autocorrelation in the subsequent sample and save storage space. Convergence was also here assessed via the potential scale reduction factor and the effective number of independent simulation draw (Gelman and Rubin 1992). As in the ECOG study, our proposed method here also needed less iterations than MCMC configuration to reach convergence and accurate results. To carry out both modeling proposals, we scaled the times to death by dividing each observation by the maximum sample value to avoid INLA numerical overflow in latency model computation. Computational times with INLA get reliable estimates in 1.24 min and the MCMC sampler needed around 7.84 min.

Table 3 shows the INLA and MCMC mean, standard deviation, and 95% credible interval of the posterior distribution of the cure proportion for allogeneic and autologous transplant patients. INLA and MCMC results are very similar.

In the case of the estimation of derived quantities of interest, we proceed analogously to the ECOG study. We estimate the INLA and MCMC posterior distribution for the cure proportion for allogeneic and autologous transplant patients (Table 4) as well as the subsequent posterior mean of the uncured survival function (Fig. 2). Outcomes also now present scarce differences and underline that allogeneic transplanted patients seem to have cure proportion levels higher than the ones for autologous patients, although we also appreciate a very broad degree of overlap.
6 Conclusions

This paper discussed an INLA approach for dealing with mixture cure models based on a general procedure by Bivand et al. (2014), Gómez-Rubio and Rue (2018), Gómez-Rubio (2018) that extends INLA to finite mixture models. We introduced latent indicators in the inferential process for classifying individuals in the cured and uncured subpopulations, and approximated the relevant posterior distribution via Gibbs sampling. In particular, we use modal Gibbs sampling (Gómez-Rubio 2018) and INLA to fit the marginal posterior distribution of each relevant element given the latent indicator variable that identifies each individual in the cure or uncured population. This is a very general and flexible proposal that allows to introduce complex modeling in both the latency and the incidence model: nonlinear covariates, temporal covariates by means of joints models, etc.

Two specific benchmark datasets from the field of medicine have been considered to illustrate our proposal. In both cases, the results support its viability and good performance, and almost entirely agree with the MCMC results. Remarkably, our proposal also shows other interesting properties such as the lower number of iterations to reach convergence and the convenient exploration of the parametric space of the latent indicators as well as faster computational times. It is worth noting that in both procedures, INLA with modal Gibbs sampling and MCMC, the specification of an appropriate number of iterations and burn-in size could be an interesting issue. In fact, the specifications in the current version of the paper have been the result of a preliminary stage where alternative possibilities have been explored, always with faster times in support of our proposal. Furthermore, the use of INLA to fit conditional models does not force the use of conjugate priors in the Gibbs sampler and avoids label switchings problems usually caused by symmetry in the likelihood function of the model parameters (Stephens 2000).

Acknowledgements Lázaro’s research was supported by a predoctoral FPU fellowship (FPU2013/02042) from the Spanish Ministry of Education, Culture and Sports. This paper was partially funded by grant MTM2016-77501-P from the Spanish Ministry of Economy and Competitiveness co-financed with FEDER funds, and by grant (SBPLY/17/180501/000491) funded by Consejería de Educación, Cultura y Deportes (JCCM, Spain) and FEDER. We wish to acknowledge two anonymous referees and the Associate Editor for their valuable comments that substantially improved the original version of the paper.

References

Akerkar R, Martino S, Rue H (2010) Implementing approximate bayesian inference for survival analysis using integrated nested laplace approximations. Prepr Stat Nor Univ Sci Technol 1:1–38
Bivand RS, Gómez-Rubio V, Rue H (2014) Approximate bayesian inference for spatial econometrics models. Spat Stat 9:146–165
Cai C, Zoua Y, Pengb Y, Zhanga J (2012) smcure: An r-package for estimating semiparametric mixture cure models. Comput Meth Prog Biomed 108:1255–1260
Christensen R, Wesley J, Branscum A, Hanson TE (2011) Bayesian ideas and data analysis: an introduction for scientists and statisticians. Chapman & Hall/CRC Press, Boca Raton
Cox DR (1972) Regression models and life-tables. J R Stat Soc: Ser B (Methodol) 34(2):187–220
Stephens M (2000) Dealing with label switching in mixture models. J R Stat Soc: Ser B (Methodol) 62(4):795–809

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.