Communications for Statistical Applications and Methods
2018, Vol. 25, No. 5, 471–488

https://doi.org/10.29220/CSAM.2018.25.5.471
Print ISSN 2287-7843 / Online ISSN 2383-4757

A Bayesian cure rate model with dispersion induced by discrete frailty

Vicente G. Cancho[a], Katherine E. C. Zavaleta[b], Márcia A. C. Macera[a],
Adriano K. Suzuki[1,a], Francisco Louzada[a]

[a]Department of Applied Mathematics and Statistics, University of São Paulo, Brazil;
[b]Department of Statistics, Federal University of São Carlos, Brazil

Abstract

In this paper, we propose extending proportional hazards frailty models to allow a discrete distribution for the frailty variable. Having zero frailty can be interpreted as being immune or cured. Thus, we develop a new survival model induced by discrete frailty with zero-inflated power series distribution, which can account for overdispersion. This proposal also allows for a realistic description of non-risk individuals, since individuals cured due to intrinsic factors (immunes) are modeled by a deterministic fraction of zero-risk while those cured due to an intervention are modeled by a random fraction. We put the proposed model in a Bayesian framework and use a Markov chain Monte Carlo algorithm for the computation of posterior distribution. A simulation study is conducted to assess the proposed model and the computation algorithm. We also discuss model selection based on pseudo-Bayes factors as well as developing case influence diagnostics for the joint posterior distribution through ψ-divergence measures. The motivating cutaneous melanoma data is analyzed for illustration purposes.

Keywords: Bayes factor, Bayesian inference, cure rate models, frailty models, Kullback-Leibler, zero-inflated power series distribution

1. Introduction

A heterogeneity between individuals should be considered when survival data comes from different groups or if individuals have repeated measurements. If heterogeneity is omitted several problems may occur such as: overestimation of relative hazard rate, biased estimates of regression coefficients, and making the regression parameters estimate tend to zero (Ata and Özel, 2013).

Frailty models are the most common models to account for unobserved heterogeneity in survival analysis (Vaupel *et al.*, 1979; Hougaard, 1986). The classical frailty model assumes a proportional hazards model conditional on the random effect (frailty) that is often a non-negative and continuous random variable. However, there are situations where a discretely-distributed frailty may be appropriate because the continuous frailty distribution does not allow for a zero-risk possibility. This may therefore represent the presence of a random number of failure per unit or the random number of causes resulting in exposure to certain damage (Moger *et al.*, 2004; Caroni *et al.*, 2010). Furthermore, discrete frailty distributions can generate units or individuals with zero frailty, describing survival models with a cure fraction. Caroni *et al.* (2010) showed that the impact of observations with zero failures may be hidden by right censoring; however, observations with a large number of failures or

[1] Corresponding author: Department of Applied Mathematics and Statistics, University of São Paulo, Avenida Trabalhador São-carlense, 400 - Centro CEP: 13566-590, São Carlos-SP, Brazil. E-mail: suzuki@icmc.usp.br

Published 30 September 2018 / journal homepage: http://csam.or.kr

causes that lead to the event will often have short lifetimes relative to baseline distribution. Therefore, a discrete frailty model can explain possible lower outliers in the data set. Another important issue is the identifiability of discrete frailty models. In the continuous framework, to achieve identifiability it is often convenient to fix the mean of the frailty variable at 1, by a suitable restriction on the frailty distribution parameters so that the scale parameter governs the variance. Thus allowing the variance to tend to zero, with unit mean, we have the no-frailty model. However, a similar problem of identifiability does not arise in the discrete frailty framework, since the frailty distribution lacks a scale parameter.

Some authors have recently developed discrete frailty models. Among these authors, Wienke (2010) considered a discrete compound Poisson process (DCPP) for frailty models, Caroni *et al.* (2010) developed discrete frailty models using Poisson, geometric and binomial distributions; and Ata and Özel (2013) proposed a discrete frailty model based on DCPP. The models proposed by these authors can be considered as particular cases of the long-term survival model proposed by Tsodikov *et al.* (2003).

Long-term survival models (cure rate models) play an important role in survival analysis where sampling units are insusceptible to the occurrence of the event of interest. The proportion of such units is usually called a cured fraction. These models were first addressed by Boag (1949) and Berkson and Gage (1952), which were known in the literature as the standard mixture model. Later, an alternative model called a promotion time cure model was proposed and investigated by Yakovlev and Tsodikov (1996) and Chen *et al.* (1999). Several other researchers have also contributed to this area such as Tsodikov *et al.* (2003), Ibrahim *et al.* (2005), Rodrigues *et al.* (2009a), Eudes *et al.* (2013), Milani *et al.* (2015), Ortega *et al.* (2015), Morita *et al.* (2016), and Cordeiro *et al.* (2016). The studies presented by these authors propose more comprehensive models. Studies on long-term survival attempt to evaluate the zeros significance in the analysis of observable and latent data. In this context, it is common to find data with possible overdispersion due to excess zeros. Then the study of this overdispersion, which can also occur due to some other causes, is important when making accurate statistical inferences. Many methods of accounting for overdispersion in various commonly used models are discussed in Morel and Neerchal (2012).

Zero-inflated models have become quite popular in the literature to deal with situations where there are excess zeros, leading to overdispersion in data (Consul and Jain, 1973; Van den Broek, 1995; del Castillo and Pérez-Casany, 2005; Yang *et al.*, 2007; Samani *et al.*, 2012; Barriga and Louzada, 2014; Choo-Wosoba *et al.*, 2015). In particular, the zero-inflated power series (ZIPS) model (Gupta *et al.*, 1995) is one method to allow for overdispersion or underdispersion. This model assumes that the sample is a mixture of two groups of individuals: one group for which the counts are generated by the standard power series distribution and another group who have zero probability of a count greater than zero. The ZIPS model is therefore a good candidate in these cases, since it includes several known zero-inflated models such as zero-inflated Poisson (ZIP), zero-inflated geometric (ZIG), zero-inflated logarithmic (ZIL), and zero-inflated negative binomial (ZINB) with overdispersion relative to the power series class models.

In this paper, we propose a new survival model induced by discrete frailty with ZIPS distribution, where zero frailty corresponds to a model containing a proportion of individuals no longer susceptible to the event of interest (long-term survivors). This model is more flexible than traditional cure rate models in regards to dispersion. Our proposal also allows a more realistic description of the non-risk individuals. Individuals cured due to intrinsic factors are modeled by a deterministic fraction; however, those cured due to an intervention are modeled by a random fraction.

We develop a Bayesian analysis for proposed model based on Markov chain Monte Carlo (MCMC) methods. To test the adequacy of the proposed model we use a Bayesian methodology based on the pseudo-Bayes factor (PSBF) (Geisser and Eddy, 1979). Another important issue is to verify the assumptions of the fitted model and perform studies to detect possible influential or extreme observations, which may cause distortions in the analysis results. These studies are called case influence diagnostics (Cook and Weisberg, 1982) and are used to help researchers interpret a model by showing the effect of data points on parameter estimates. Thus, we believe that the development of case influence diagnostic Bayesian tools for the model proposed in this paper are a significant contribution. Our goal is to also develop diagnostic measures from a Bayesian point of view based on the ψ-divergence (Peng and Dey, 1995) between the posterior distributions of the parameters of the proposed model.

Our model is illustrated with a dataset on skin cancer. The data are part of a study on cutaneous melanoma for the evaluation of post-operative treatment performance as a method to prevent recurrence. The data suggest the existence of a fraction of individuals with zero-risk (no longer susceptible or cured). It is therefore of interest to estimate this zero-risk fraction; in addition, it would also be interesting to obtain the proportion of individuals who are cured due to genetic factors or other inherent characteristics.

The paper is organized as follows. In Section 2 we give the survival functions for the frailty model with discrete distributions. The proposed model formulation is presented in Section 3. In Section 4 we describe Bayesian computational strategies using a MCMC algorithm and discuss some measures of model selection as well as Bayesian case diagnostics based on ψ-divergence. The simulation study in Section 5 evaluates the performance of the proposed method. In Section 6, a real dataset on cutaneous melanoma is used to illustrate the applicability of the methodology. Finally, the article is concluded with a further discussion of the proposed model in Section 7.

2. Survival function for discrete frailty model

Frailty models provide a suitable method to consider the existence of possible associations and unobserved heterogeneity in survival models. A standard frailty model is often built on a proportional hazards framework conditional on a non-negative random variable (frailty) Z. The effect of a frailty Z is to modify a baseline hazard function $h_0(t)$ to $Zh_0(t)$ for an individual. The frailty model can then be written in terms of its conditional survival function

$$S(t|Z) = \Pr(T > t|Z) = \exp\left(-Z\int_0^t h_0(u)du\right) = \exp\left(-ZH_0(t)\right), \tag{2.1}$$

where $H_0(t) = \int_0^t h_0(u)du$ is the baseline cumulative hazard function. The marginal survival function, $S(t)$, can be obtained by integrating the conditional survival function in Equation (2.1) over the domain of the distribution of Z, once the frailty distribution is specified. So, the marginal survival function is given by

$$S(t) = \int S(t|z)dF(z) = E(S(t|Z)) = \mathcal{L}[H_0(t)], \tag{2.2}$$

where $F(z)$ is the distribution function of Z and $\mathcal{L}(s)$ is the Laplace transform of Z.

In Equation (2.1), if the frailty Z can take non-negative integer values, i.e., Z has a discrete distribution on $\{0, 1, 2, \dots\}$ specified by $\Pr[Z = z] = p_z$ for $z = 0, 1, 2, \dots$, then, assuming a proportional hazards model, the marginal survival function can be written as

$$S(t) = \sum_{z=0}^{\infty} S(t|z)p_z = E_Z(S(t|Z)) = G_Z[S_0(t)], \tag{2.3}$$

where G_Z is the probability generating function (pgf) of Z and $S_0(t) = \exp(-H_0(t))$ is the baseline survival function. Note that $S(t)$ given in Equation (2.3) is the survival function proposed by Rodrigues *et al.* (2009b). Moreover, if $\Pr(Z = 0) > 0$ the survival function in (2.3) is an improper function, i.e., $\lim_{t\to\infty} S(t) = G_Z(0) = \Pr(Z = 0) > 0$. This can be taken to describe survival models with a cure fraction. In case that $\Pr(Z = 0) = 0$, the survival function is proper, i.e., $\lim_{t\to\infty} S(t) = G_Z(0) = \Pr(Z = 0) = 0$ and $\lim_{t\to 0} S(t) = G_Z(1) = 1$.

Discrete distributions such as Poisson, binomial, geometric or negative binomial have recently been considered for the frailty models (Caroni *et al.*, 2010; Ata and Özel, 2013). In this study, we derive the survival function of the frailty model using ZIPS distribution with probability function given by

$$P(Z;\theta,\phi) = \begin{cases} \dfrac{\phi + a_0 A(\theta)^{-1}}{1+\phi}, & z = 0, \\ \dfrac{a_z\theta^z A(\theta)^{-1}}{1+\phi}, & z = 1,2,\ldots, \end{cases} \tag{2.4}$$

where $a_z > 0$ depends only in z, $A(\theta) = \sum_{z=1}^{\infty} a_z\theta^z$, $\theta \in (0, s)$ (s can be ∞) is such that $A(\theta)$ is finite and $-a_0/A(\theta) \le \phi < \infty$. For more details on the ZIPS class of distributions, see Gupta *et al.* (1995) and Van den Broek (1995). When $\phi = 0$, the ZIPS model reduces to a standard power series model. However, the case $\phi > 0$ indicates overdispersion.

3. The proposed model

In this section, we derive the survival function of our frailty model with a cure fraction using a ZIPS distribution as given in Equation (2.4). To obtain the unconditional survival, let us define Z as a ZIPS-distributed random variable, where the p.g.f is given by

$$G_Z(\xi) = \frac{\phi}{1+\phi} + \frac{A(\theta\xi)}{(1+\phi)A(\theta)}, \tag{3.1}$$

which converges when $|\xi| \le 1$. Using Equation (2.3) together with (3.1), the unconditional survival function is obtained as

$$S(t) = \frac{\phi + A(\theta S_0(t))A(\theta)^{-1}}{1+\phi}, \quad t > 0. \tag{3.2}$$

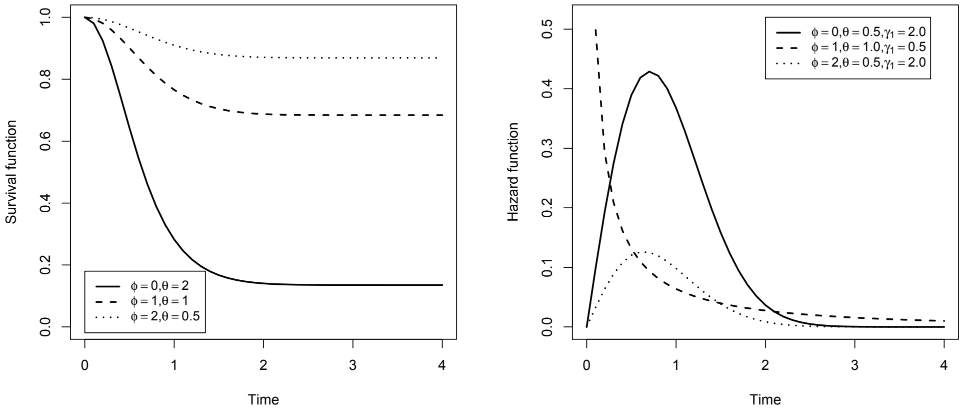
Since $\lim_{t\to\infty} S(t) = [\phi/(1+\phi)] + [a_0/(1+\phi)A(\theta)] > 0$, the survival function in Equation (3.2) is an improper function. Note that if $\phi = 0$ in (3.2), then $S(t) = A(\theta S_0(t))/A(\theta)$, which results in the model proposed by Cancho *et al.* (2013). This model includes as particular cases the models proposed by Caroni *et al.* (2010).

We refer to the model in Equation (3.2) as the ZIPS survival model with a cure fraction, or simply the ZIPS-CF model.

Using Equation (3.2) we can also obtain the proportion of zero-risk (cured fraction), which is given by

$$p_0 = \lim_{t\to\infty} S(t) = \frac{\phi}{1+\phi} + \frac{a_0}{(1+\phi)A(\theta)}. \tag{3.3}$$

It can be observed that, since $\theta \in (0, s)$ (s can be ∞), when $\theta \to s$ the proportion of zero-risk tends to deterministic zero, $\phi/(1+\phi)$, while $\theta \to 0$ the proportion of zero-risk tends to 1.



(a) (b)

Figure 1: *(a) Survival functions of the ZIPS-CF model for* $\gamma_1 = 2$, $\gamma_2 = 1$ *and different values of* ϕ *and* θ, *and (b) Hazard functions of the ZIPS-CF model for* $\gamma_2 = 1$ *and different values of* ϕ, θ *and* γ_1. *ZIPS-CF = zero-inflated power series survival model with a cure fraction.*

The proportion of zero-risk given in Equation (3.3) describe individuals who are not at risk because they did not experience the event of interest. In our model, the cured fraction p_0 has two components. The first one is a deterministic fraction of zero-risk referred to immune subjects, i.e., those patients who are cured due to inherent characteristics (e.g., genetic characteristics or immune resistance). The second component is a random fraction of zero-risk referred to individuals who were initially at risk and by an intervention (e.g., treatment) are cured. Thus, the decomposition of this fraction is important since it allows determining the proportion of patients cured due to genetic factors and the proportion cured due to an intervention. The cured fraction by an intervention is then given by

$$p_0^* = \frac{a_0}{(1+\phi)A(\theta)}. \tag{3.4}$$

The corresponding probability density function (pdf) of the model (3.2) is given by the expression

$$f(t) = -S'(t) = \frac{\theta f_0(t) A'(\theta S_0(t))}{(1+\phi)A(\theta)}, \tag{3.5}$$

and the hazard function is

$$h(t) = \frac{\theta f_0(t) A'(\theta S_0(t)) A(\theta)^{-1}}{\phi + A(\theta S_0(t)) A(\theta)^{-1}}, \tag{3.6}$$

where $A'(\theta S(t)) = A'(\theta) \mid_{\theta = \theta S_0(t)}$ and $f_0(t) = -dS_0(t)/dt$ is a baseline density function. The functions $f(t)$ and $h(t)$ are improper, since the survival function $S(t)$ is an improper function.

Figure 1 illustrates some possible shapes of the survival and hazard functions of the ZIPS-CF model for some selected values of ϕ and θ. For the illustration, we assume a Weibull distribution as the baseline distribution with γ_1 and γ_2 the shape and scale parameters, respectively.

The survival function of the individuals susceptible to the event of interest (or at-risk individuals), denoted by $S_{\text{nc}}(t)$, can be expressed as

$$S_{\text{nc}}(t) = P(T > t | Z \geq 1) = \frac{A(\theta S_0(t)) - a_0}{A(\theta) - a_0}. \tag{3.7}$$

Table 1: Long-term survival function ($S(t)$), hazard function ($h(t)$) and cured fraction (p_0) for different special cases

Model	$S(t)$	$h(t)$	p_0
ZIP-CF	$\frac{\phi+e^{-\theta(1-S_0(t))}}{(1+\phi)}$	$\frac{\theta f_0(t)e^{-\theta(1-S_0(t))}}{\phi+e^{-\theta(1-S_0(t))}}$	$\frac{\phi}{(1+\phi)}+\frac{e^{-\theta}}{(1+\phi)}$
ZIG-CF	$\frac{\phi}{1+\phi}+\frac{(1-\theta)}{(1+\phi)(1-\theta S_0(t))}$	$\frac{\theta(1-\theta)f_0(t)(1-\theta S_0(t))^2}{\phi+(1-\theta)(1-\theta S_0(t))^{-1}}$	$\frac{\phi}{1+\phi}+\frac{(1-\theta)}{(1+\phi)}$
ZIL-CF	$\frac{\phi}{1+\phi}+\frac{\log(1-\theta S_0(t))}{(1+\phi)S_0(t)\log(1-\theta)}$	$\frac{\frac{-f_0(t)}{S_0(t)^2\log(1-\theta)}\left(\frac{\theta S_0(t)}{1-S_0(t)}+\log(1-\theta S_0(t))\right)}{\phi+\frac{\log(1-\theta S_0(t))}{S_0(t)\log(1-\theta)}}$	$\frac{\phi}{1+\phi}-\frac{\theta}{(1+\phi)\log[1-\theta]}$

ZIP-CF = zero-inflated Poisson model with a cure fraction; ZIG-CF = zero-inflated geometric model with a cure fraction; ZIL-CF = zero-inflated logarithmic model with a cure fraction.

It can be noted that $S_{\text{nc}}(0) = 1$ and $S_{\text{nc}}(\infty) = 0$, i.e., S_{nc} is a proper survival function. We can obtain several recently proposed classes of distributions from the model given in (3.7), considering different choices for the distribution of the frailty variable Z and for the distribution of baseline survival $S_0(t)$. For example, the Weibull-geometric class proposed by Barreto-Souza *et al.* (2011) is obtained, considering that Z has a geometric distribution and the baseline distribution is a Weibull. However, if the baseline distribution is exponential with rate $\theta > 0$ then we obtain the distribution introduced by Chahkandi and Ganjali (2009). Members of this distribution class have a decreasing failure rate that include classes of distributions proposed by Adamidis and Loukas (1998), Coskun (2007), and Tahmasbi and Rezaei (2008).

There is a mathematical relationship between the ZIPS-CF model and the standard mixture model (Boag, 1949; Berkson and Gage, 1952), which can be written as

$$S(t) = p_0 + (1 - p_0)S_{\text{nc}}(t), \tag{3.8}$$

where p_0 is given by Equation (3.3) and the uncured fraction is then given by $1-p_0 = (1-a_0A(\theta)^{-1})(1+\phi)^{-1}$. Based on results of Li *et al.* (2001), the ZIPS-CF models represented as mixture model, Equation (3.8), are identifiable as long as $S_{\text{nc}}(t)$ is a parametrically specified function, which in this case is given in Equation (3.7).

3.1. Special cases

The ZIPS class includes several important and known zero-inflated distributions, such as zero-inflated binomial (ZIB), ZIP, ZIG, ZIL, and ZINB. In this section we only present three special cases of the ZIPS-CF model, particularly when the frailty variable Z assumes ZIP, ZIG, and ZIL distributions. For example, if $a_z = 1/z!$ and $A(\theta) = e^{\theta}$, $\theta > 0$, then Equation (2.4) defines the ZIP distribution. By taking $a_z = 1$ and $A(\theta) = (1-\theta)^{-1}$, $\theta \in (0, 1)$ we obtain the ZIG distribution. The ZIL distribution is obtained by considering $a_z = 1/(z + 1)$ and $A(\theta) = -\log(1 - \theta)/\theta$ with $\theta \in (0, 1)$ in (2.4). Table 1 presents expressions for long-term survival and improper hazard functions as well as for the proportion of zero-risk to the specific models by considering these distributions. Note that if $\phi = 0$ in the ZIP-CF model, we obtain the promotion time cure model introduced by Chen *et al.* (1999).

4. Bayesian computational strategies

Let us consider $T_1, \ldots, T_n$ the lifetime of n individuals. Suppose the lifetime is not completely observed and it is subject to right censoring, where C_i denote censoring time. We then observe $t_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \leq C_i)$, where $\delta_i = 1$ if T_i is a lifetime and $\delta_i = 0$ if it is right censored, for

$i = 1, \dots, n$. We incorporate the covariates through θ. For each individual i, let $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$ denote the vector of covariates, and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ denote the corresponding vector of regression coefficients. We relate θ to the covariates by $g(\theta_i) = \eta(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $i = 1, \dots, n$. A possible link function that can be adopted for the ZIP-CF model is the logarithmic link function, given by $g(\theta_i) = \log(\theta_i)$, while for the ZIG-CF and ZIL-CF models we can adopt the logistic link function, given by $g(\theta_i) = \log\{\theta_i/(1-\theta_i)\}$. In all cases, the models are identifiable in the sense of Li *et al.* (2001).

From n independent individuals, the likelihood function under non-informative censoring is given by

$$L(\mathcal{D}|\boldsymbol{\vartheta}) = \frac{1}{(1+\phi)^n} \prod_{i=1}^{n} \frac{1}{A(\theta_i)} \left[\theta_i f_0(t_i;\gamma) A'(\theta_i S_0(t_i;\gamma))\right]^{\delta_i} \left[\phi A(\theta_i) + A(\theta_i S_0(t))\right]^{1-\delta_i}, \tag{4.1}$$

where $\boldsymbol{\vartheta} = (\phi, \boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top)^\top$ and $\mathcal{D} = (\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{x})$, whereas $\boldsymbol{t} = (t_1, \dots, t_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^\top$, and $\boldsymbol{x} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})^\top$. We now assume a Weibull distribution as a baseline distribution with $h_0(t) = \gamma_1 \gamma_2 t^{\gamma_1 - 1}$, for $\gamma_1 > 0$ and $\gamma_2 > 0$. By assuming this baseline distribution, we allow a greater flexibility to the hazard function of the model.

Inferential methods are investigated from a Bayesian point of view. In this context, a prior distribution for common parameters are required; in addition, we also adopt proper prior distributions for all model parameters to ensure proper posterior distributions. As is typically carried out for regression parameters, normal priors are specified for β_j. Gamma priors are specified for baseline distribution parameters. A uniform prior that depends on $A(\theta_i)$ is adopted for ϕ, the overdispersion parameter. In summary, the priors chosen for our analysis are as follows

$$\begin{aligned}
&\phi|\boldsymbol{\beta} \sim U(\max(-1,\omega), \rho), \quad \omega = \max\left\{-\frac{a_0}{A(\theta_i)}\right\}, \ \rho > 0,\\
&\beta_j \sim N\left(\mu_j, \sigma_j^2\right), \quad j = 0, 1, \dots, p,\\
&\gamma_k \sim \text{Gamma}(a_k, b_k), \quad k = 1, 2,
\end{aligned}$$

where the parameters ρ, μ_j, σ_j^2, a_k, and b_k are hyperprior parameters, $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, Gamma(a, b) denotes Gamma distribution with a as shape and b as scale (and mean a/b), and $U(c, d)$ denotes a uniform distribution on (c, d). Then, the joint prior density of all parameters is obtained by

$$\pi(\boldsymbol{\vartheta}) = \pi(\phi|\boldsymbol{\beta})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}) \propto \frac{\gamma_1^{a_1-1}\gamma_2^{a_2-1}\exp\left(-\sum_{k=1}^{2} b_k\gamma_k - \sum_{j=0}^{p}\frac{(\beta_j-\mu_j)^2}{2\sigma_j^2}\right)}{\rho - \max(-1,\omega)}. \tag{4.2}$$

Combining the likelihood function in Equation (4.1) with the prior distribution in (4.2), the joint posterior distribution for $\boldsymbol{\vartheta}$ is obtained as

$$\pi(\boldsymbol{\vartheta}|\mathcal{D}) \propto \frac{1}{(1+\phi)^n} \prod_{i=1}^{n} \frac{1}{A(\theta_i)} \left[\theta_i f_0(t_i;\gamma) A'(\theta_i S_0(t_i;\gamma))\right]^{\delta_i} \left[\phi A(\theta_i) + A(\theta_i S_0(t))\right]^{1-\delta_i} \pi(\boldsymbol{\vartheta}). \tag{4.3}$$

The joint posterior density in Equation (4.3) is analytically intractable, so we based our inference on the MCMC simulation methods such as the Gibbs sampler (Casella and George, 1992) and Metropolis-Hastings (Chib and Greenberg, 1995). To obtain a posterior sample of ϕ, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$, the Gibbs sampler has proved to be a powerful tool. Thus, we first obtain the full conditional distributions of the parameters given as

$$\begin{aligned}
\pi(\phi|\boldsymbol{\beta},\boldsymbol{\gamma},\mathcal{D}) &\propto \frac{1}{(1+\phi)^n}\prod_{i=1}^{n}\left[\phi A(\theta_i)+A(\theta_i S_0(t_i;\gamma))\right]^{1-\delta_i}\ \pi(\phi|\boldsymbol{\beta}),\\
\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\phi,\mathcal{D}) &\propto \prod_{i=1}^{n}\frac{1}{A(\theta_i)}\left[\theta_i A'(\theta_i S_0(t_i;\gamma))\right]^{\delta_i}\left[\phi A(\theta_i)+A(\theta_i S_0(t_i;\gamma))\right]^{1-\delta_i}\ \pi(\boldsymbol{\beta}), \qquad (4.4)\\
\pi(\boldsymbol{\gamma}|\boldsymbol{\beta},\phi,\mathcal{D}) &\propto \prod_{i=1}^{n}\left[f_0(t_i;\gamma)A'(\theta_i S_0(t_i;\gamma))\right]^{\delta_i}\left[\phi A(\theta_i)+A(\theta_i S_0(t_i;\gamma))\right]^{1-\delta_i}\ \pi(\gamma).
\end{aligned}$$

Note that the distributions given in (4.4) are not standard distributions, and we need to perform Metropolis-Hastings steps within the Gibbs sampler to simulate samples of $\boldsymbol{\vartheta}$. MCMC computations were implemented using the JAGS from R system (R Development Core Team, 2010). The R codes can be obtained by e-mail from the first author.

A comparison of models for a given data set and selecting the model that best fits the data is an important issue in data analysis. Our interest is to test the adequacy of the ZIPS-CF model through the zero-risk parameter for data modeled by the fraction of deterministic risk. In particular, we are interested in testing the hypotheses $H_0 : \phi = \phi_0$ versus $H_1 : \phi \neq \phi_0$, where ϕ_0 is a specified value for ϕ. When $\phi_0 = 0$, the hypotheses allows us to evaluate the adequacy of the survival model proposed by Cancho *et al.* (2013). There are various methodologies to test above hypotheses, and a Bayesian alternative is by determining the PSBF of the model under the restricted hypothesis (M_0 model) with respect to the unrestricted hypothesis (M_1 model). Following Geisser and Eddy (1979), in this paper we use the PSBF based on the marginal predictive likelihoods of each model, M_0 and M_1, which is obtained as a product of the conditional predictive ordinates (CPO) (Gelfand *et al.*, 1992) given by

$$f\left(y_r|\mathcal{D}^{(-r)}\right) = \int_{\Theta} f(y_r|\boldsymbol{\vartheta})\pi\left(\boldsymbol{\vartheta}|\mathcal{D}^{(-r)}\right)d\boldsymbol{\vartheta}, \quad r = 1,\ldots,n, \qquad (4.5)$$

where $\mathcal{D}^{(-r)}$ denotes data with the rth observation deleted. Let $f(y_r|\mathcal{D}^{(-r)}, M_j)$, $j = 0, 1$, denote the CPO$_r$ for the rth observation of the model under H_j, $j = 0, 1$. The PSBF of M_0 with respect to M_1 can be expressed as

$$\text{PSBF}_{01} = \frac{\prod_{r=1}^{n} f\left(y_r|\mathcal{D}^{(-r)}, M_0\right)}{\prod_{r=1}^{n} f\left(y_r|\mathcal{D}^{(-r)}, M_1\right)}. \qquad (4.6)$$

Then, PSBF$_{01}$ < 1 provide evidence against of M_0 model. Thus, to compare the ZIPS-CF model to the power series model with a cure fraction proposed by Cancho *et al.* (2013), i.e., $H_0 : \phi = 0$ versus $H_1 : \phi \neq 0$, we can calculate an estimate of the CPO by Monte Carlo integration with posterior samples. Let $\boldsymbol{\vartheta}^{(1)},\ldots,\boldsymbol{\vartheta}^{(S)}$ be a posterior sample of the model parameters under H_0, an estimate of the CPO for the M_0 model (power series model with a cure fraction) can be expressed as

$$\hat{f}\left(y_r|\mathcal{D}^{(-r)}, M_0\right) = \begin{cases} \dfrac{1}{S}\displaystyle\sum_{i=1}^{S}\left(\theta^{(i)} f_0(t_r)\frac{A'\left(\theta^{(i)}S_0(t_r)\right)}{A\left(\theta^{(i)}\right)}\right)^{-1}, & \text{if } \delta_r = 1,\\[2ex] \dfrac{1}{S}\displaystyle\sum_{i=1}^{S}\left(\frac{A\left(\theta^{(i)}S_0(t_r)\right)}{A\left(\theta^{(i)}\right)}\right)^{-1}, & \text{if } \delta_r = 0, \end{cases} \qquad (4.7)$$

where δ_r indicates whether the rth observation is a complete observation ($\delta_r = 1$) or right censored ($\delta_r = 0$).

4.1. Case influence analysis

A common diagnostic tool to assess the influence of an observation on the fit of a model is the case influence (or deletion) diagnostic (Cook and Weisberg, 1982). Case influence diagnostic is useful for case deletion, outlier detection, or model modification. In addition, it helps researchers interpret a model by showing the effect of data points on parameter estimates. Let $D_\psi(P, P_{(-i)})$ denote the measure of divergence ψ between P and $P_{(-i)}$, where P is the full-data posterior distribution (given all observations) and $P_{(-i)}$ is the case-deleted posterior distribution (excluding case i). $D_\psi(P, P_{(-i)})$ measures the effect of deleting the ith observation from the full data on the posterior distribution of $\boldsymbol{\vartheta}$. This measure can be expressed as (with respect to Lebesgue measure)

$$D_\psi(P, P_{(-i)}) = \int \pi(\boldsymbol{\vartheta}|\mathcal{D})\, \psi\left(\frac{\pi(\boldsymbol{\vartheta}|\mathcal{D})}{\pi(\boldsymbol{\vartheta}|\mathcal{D}^{(-i)})}\right) d\boldsymbol{\vartheta}. \tag{4.8}$$

$D(P, P_{(-i)}) \neq D(P_{(-i)}, P)$ and ψ is often a convex function of $\psi(1) = 0$. Dey and Birmiwal (1994) presents several possible ways for the ψ, e.g., if $\psi(u) = -\log(u)$ we get the Kullback-Leibler (KL) divergence; if $\psi(u) = (u - 1)\log(u)$ then we have the J-divergence, which is a symmetric version of the KL divergence; the variational distance or L_1-distance is obtained by $\psi(u) = 0,5|u - 1|$; and finally, if $\psi(u) = (u - 1)^2/u$ we have the χ^2-distance.

There is a mathematical relationship between the CPO defined in Equation (4.5) and the divergence measure ψ, which can be written as

$$D_\psi\left(P, P_{(-i)}\right) = E_{\boldsymbol{\vartheta}|\mathcal{D}}\left[\psi\left(\frac{\text{CPO}_i}{L(\mathcal{D}_i|\boldsymbol{\vartheta})}\right)\right], \tag{4.9}$$

where $E_{\boldsymbol{\vartheta}|\mathcal{D}}(\cdot)$ denotes the expected value with respect to joint posterior distribution $\pi(\boldsymbol{\vartheta}|\mathcal{D})$ and $L(\mathcal{D}_i|\boldsymbol{\vartheta})$ is the likelihood function of the ith observation. It therefore follows from Equation (4.9) that KL divergence (Cho *et al.*, 2009) can be represented by

$$D_{\text{KL}}\left(P, P_{(-i)}\right) = -\log(\text{CPO}_i) + E_{\boldsymbol{\vartheta}|\mathcal{D}}\{\log L(\mathcal{D}_i|\boldsymbol{\vartheta})\}. \tag{4.10}$$

An estimate of (4.10) can be obtained generating samples of the posterior distribution of $\boldsymbol{\vartheta}$ via MCMC. Let $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(S)}$ be a sample of size S of $\pi(\boldsymbol{\vartheta}|\mathcal{D})$, an estimate of the KL divergence is given by

$$D_{\text{KL}}\widehat{(P, P_{(-i)})} = -\log\left(\widehat{\text{CPO}}_i\right) + \frac{1}{S}\sum_{s=1}^{S} \log L\left(\mathcal{D}_i|\boldsymbol{\vartheta}^{(s)}\right). \tag{4.11}$$

According to Peng and Dey (1995) and Weiss (1996), the ith observation is considered influential when $D_{L_1} > 0.30$ or $D_{\chi^2} > 0.56$. Thus, if we use the J-divergence, an observation which $D_J > 0.42$ can be considered as influential. Similarly, using the KL divergence, we can consider an influential observation when $D_{\text{KL}} > 0.22$.

5. Simulation study

In this section, we conduct a simulation study to evaluate the frequentist properties of Bayesian estimates of ZIPS-CF model parameters as well as verify the performance of PSBF to discriminate between two models. For simplicity, the study was based on the generation of a data set from the ZIP-CF

Table 2: Posterior summaries for the parameters of the ZIP-CF model based on 1,000 simulated data sets

n		Scenario 1					Scenario 2				
		ϕ	γ_1	γ_2	β_0	β_1	ϕ	γ_1	γ_2	β_0	β_1
100	True values	0.50	2.00	2.00	2.00	−2.00	2.00	2.00	2.00	2.00	−2.00
	post. mean	0.55	2.09	2.25	2.21	−1.96	2.27	2.22	2.49	2.29	−2.02
	post. sd.	0.17	0.25	1.02	0.60	0.48	0.65	0.39	0.97	0.74	0.76
	bias	0.05	0.09	0.25	0.21	0.04	0.27	0.22	0.49	0.29	−0.02
	RMSE	0.18	0.27	1.05	0.63	0.48	0.70	0.45	1.08	0.80	0.76
	CP	0.97	0.93	0.96	0.97	0.94	0.96	0.91	0.99	0.98	0.94
200	True values	0.50	2.00	2.00	2.00	−2.00	2.00	2.00	2.00	2.00	−2.00
	post. mean	0.54	2.05	2.16	2.12	−1.99	2.19	2.08	2.23	2.22	−1.96
	post. sd.	0.12	0.17	0.78	0.40	0.32	0.47	0.26	0.96	0.58	0.53
	bias	0.04	0.05	0.16	0.12	0.01	0.19	0.08	0.23	0.22	0.04
	RMSE	0.13	0.18	0.79	0.42	0.32	0.50	0.27	0.99	0.62	0.53
	CP	0.95	0.94	0.94	0.97	0.94	0.97	0.94	0.98	0.98	0.93
400	True values	0.50	2.00	2.00	2.00	−2.00	2.00	2.00	2.00	2.00	−2.00
	post. mean	0.53	2.02	2.06	2.06	−1.99	2.13	2.05	2.03	2.22	−1.96
	post. sd.	0.08	0.12	0.51	0.25	0.23	0.35	0.17	0.78	0.44	0.33
	bias	0.03	0.02	0.06	0.06	0.01	0.13	0.05	0.03	0.22	0.04
	RMSE	0.09	0.13	0.51	0.26	0.23	0.38	0.17	0.78	0.49	0.33
	CP	0.97	0.94	0.95	0.97	0.92	0.95	0.95	0.92	0.97	0.94

ZIP-CF = zero-inflated Poisson model with a cure fraction; post. mean = posterior mean of the parameter estimates; post. sd. = the posterior standard deviation; RMSE = root mean squared error; CP = coverage probabilities.

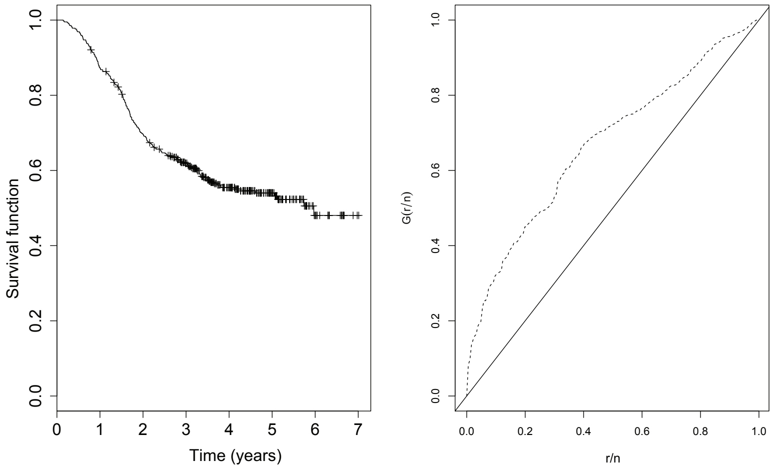
model with a Weibull baseline distribution, for which the hazard function is given by $h_0(t) = \gamma_1\gamma_2 t^{\gamma_1 - 1}$, for different sample sizes. In the simulation we fixed $\gamma_1 = 2$, $\gamma_2 = 2$ and considered two scenarios for the overdispersion parameter (1) $\phi = 0.5$ and (2) $\phi = 2.0$. We consider that parameter θ is related to the covariate x through the logarithmic function, i.e., $\log(\theta_i) = \beta_0 + \beta_1 x_i$, $i = 1, \ldots, n$, with $\beta_0 = 2$ and $\beta_1 = -2$. The true value parameter for $\gamma_1, \gamma_2, \beta_0$, and β_1 were selected arbitrarily. The covariate x_i is a binary covariate taking value 0 or 1 with probability 0.5. For the simulation studies, we generated the survival data with about 65% censored observations using a distribution $U(0, \tau)$, where τ controls the percentage of censored observations. We simulated 1,000 data sets from the ZIP-CF model, each of size $n = 100$, $n = 200$, and $n = 400$.

To express noninformative prior distributions in (4.2), we specify $\gamma_k \sim \text{Gamma}(1, 0.01)$, $k = 1, 2$; $\beta_j \sim N(0, 100)$, $j = 0, 1$ and $\phi|\boldsymbol{\beta} \sim U(\max(-1, \omega), 100)$, where $\omega = \max\{-\exp(\theta_i)\}$. For each simulated data set, the MCMC algorithm run was based on two chains of 25,000 iterations for each one. The first 5,000 iterations were discarded as a burn-in period to eliminate the effect of the initial values. We took a spacing of size 10 to avoid the correlation between the generated values that resulted in a sample of size 1,000 for each chain. The convergence of the chains was monitored using the methods of Cowles and Carlin (1996). Table 2 shows the summarized results after the burn-in period, based on 1,000 replicates. Each parameter of the ZIP-CF model are presented in the posterior mean of the parameter estimates (post. mean), posterior standard deviation (post. sd.), root mean squared error (RMSE), and the estimate of bias and the coverage probabilities (CP) of the 95% highest posterior density interval (HPD interval). Similar results are held by other models. The simulation study suggests that the fitted ZIP-CF model provides accurate estimates for all model parameters. We see that the point posterior estimates of the parameters are close to their true values. The estimates of bias and RMSE decrease as the sample sizes increases. Furthermore, the coverage probabilities are close to the nominal coverage level of 95%.

We have also conducted a simulation study based on the PSBF to test hypotheses $H_0 : \phi = 0$ versus $H_1 : \phi \neq 0$ in order to evaluate the adequacy of the ZIP-CF model. These hypotheses are equivalent

Table 3: Percentage of times that $H_0 : \phi = 0$ is rejected based on 1,000 simulations

ϕ	n			
	100	200	300	400
0.0	12.1	11.2	10.9	10.2
0.5	84.4	98.8	99.0	100.0
2.0	85.2	97.0	100.0	100.0



(a) Kaplan-Meier (b) TTT-plot

Figure 2: *Kaplan-Meier curve and total time on test plot (TTT-plot) for cutaneous melanoma data.*

to the ZIP-CF model (M_1 model) and the promotion time cure model proposed by Chen *et al.* (1999) (M_0 model). Samples of different sizes ($n = 100$, $n = 200$, $n = 300$, and $n = 400$) of the ZIP-CF model were generated with the same parameters β_0, β_1, γ_1, and γ_2 given above and different values for the overdispersion parameter, $\phi \in \{0, 0.5, 2.0\}$. We generated 1,000 samples for each configuration. Table 3 presents the percentage of times the null hypothesis is rejected considering the PSBFs of the M_0 model with respect to the M_1 model; therefore, the null hypothesis is rejected when $\text{PSBF}_{01} < 1$. The results indicate that the PSBF can discriminate the model adequately and satisfactorily.

6. Application: the cutaneous melanoma data

In this section we demonstrate an application of the proposed model described in Section 3 and the estimation procedure by using a real data set collected in a study on cutaneous melanoma in order to evaluate the treatment performance with a high dose of a certain drug (interferon alfa-2b) as a method to prevent recurrence. The sample included 417 patients without missing values who entered in the study from 1991 to 1995, and were followed up until 1998. This data set was analyzed and published by Kirkwood *et al.* (2000). The data provide the time (in years) until patient death or the censoring time (mean = 3.18 and sd = 1.69), with 56% of censored observations. The covariates that compose the data set are: treatment (x_1) (0: without treatment, $n = 204$; 1: interferon, $n = 213$); age (x_2) (in years; mean = 48.0 and sd. = 13.1); nodule category (x_3) (1: $n = 82$; 2: $n = 87$; 3: $n = 137$; 4: $n = 111$); sex (x_4) (0: male, $n = 263$; 1: female, $n = 154$); functional capacity (x_5) (0: active, $n = 363$; 1: others, $n = 54$) and tumor thickness (x_6) (in mm, mean = 3.94 and sd. = 3.20).

Figure 2 presents the Kaplan-Meier estimate of the survival function and the total time on test plot

Table 4: Bayesian criteria for the two fitted models, namely ZIP-CF and ZIG-CF

Criteria	Model			
	ZIP-CF	ZIG-CF	P-CF	G-CF
DIC	1032.037	1029.056	1038.546	1045.527
EAIC	1042.295	1039.814	1047.360	1057.103
EBIC	1082.625	1080.145	1083.658	1093.401

ZIP-CF = zero-inflated Poisson model with a cure fraction; ZIG-CF = zero-inflated geometric model with a cure fraction; DIC = deviance information criterion; EAIC = expected Akaike information criterion; EBIC = expected Bayesian information criterion.

(TTT-plot) (Sun and Kececloglu, 1999) for the melanoma data. The behavior of the Kaplan-Meier curve in Figure 2(a) clearly suggests the existence of a fraction of individuals with zero-risk, thus a model that ignores the possibility of a cure fraction is unsuitable to analyze these data. Moreover, the TTT-plot in Figure 2(b) indicates an increasing hazard function, so that the Weibull model can be an adequate alternative for baseline distribution.

Following Barral (2001), cutaneous melanoma is a common cancer of the skin whose incidence is dramatically increasing in persons with light-colored skin in all parts of the world. This disease is resistant to traditional chemotherapy and radiotherapy; therefore, malignant melanoma is a favorite target for alternative therapies that often involve immunological mechanisms. Taking into account this fact and the conclusions from the analysis of Figure 2, it is reasonable to assume that a proportion of the patients are cured (no longer susceptible) and it is also of interest to estimate the fraction of individuals cured due to treatment or therapy. Thus we applied the proposed ZIPS-CF model with a Weibull baseline distribution to cutaneous melanoma data. For our purposes, we link the parameter θ to the six covariates described above: treatment, age, nodule category, sex, functional capacity, and tumor thickness. The linking between the parameter θ and the covariates is therefore expressed through

$$\log(\theta_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \ \text{(ZIP-CF)} \quad \text{or} \quad \log\left(\frac{\theta_i}{1-\theta_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \ \text{(ZIG-CF)}, \tag{6.1}$$

with $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})^\top$, $i = 1, \dots, n$, and $\boldsymbol{\beta} = (\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6)^\top$.

Similar to the simulation study, the priors are specified as follows. Normal priors were specified for regression parameters: $\beta_j \sim N(0, 100)$, $(j = 0, \dots, 6)$; Gamma priors were adopted for baseline distribution parameters: $\gamma_k \sim \text{Gamma}(1, 0.01)$, $(k = 1, 2)$; a uniform prior on $(\max(-1, \omega), 100)$ was adopted for $\phi : \phi \sim U(\max(-1, \omega), 100)$, $(\omega = \max\{-e^{\theta_i}\}$ for ZIP-CF model, $\omega = \max\{-(1-\theta_i)^{-1}\}$ for ZIG-CF model). The joint posterior density is then obtained through Equation (4.3). In order to obtain the posterior samples, the MCMC run was based on based on two chains of 25,000 iterations for each one, where the first 5,000 iterations were discarded as a burn-in period.

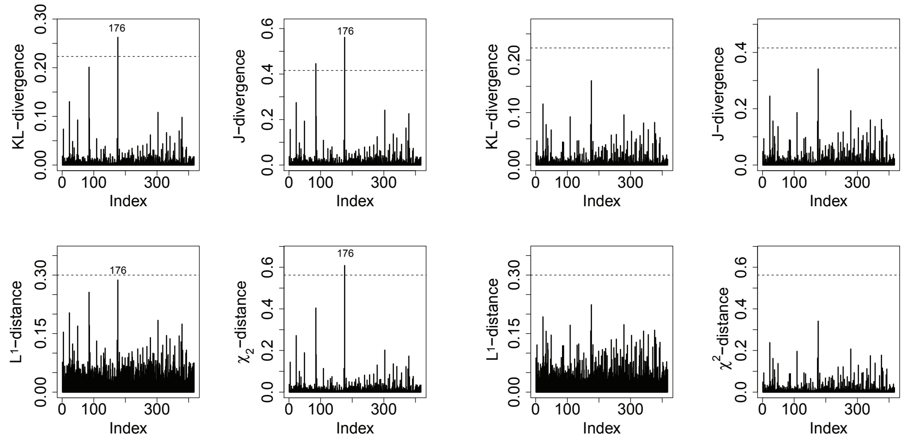
Model comparison can be performed considering the PSBF. The Monte Carlo estimate of the PSBF of ZIG-CF model (M_0 model) with respect to ZIP-CF model (M_1 model) yielded $\text{PSBF}_{01} = 4.5$ indicates strong evidence in favor of a ZIG-CF model. In order to compare the models, and also with the alternative models: Poisson model with a cure fraction (P-CF model) and Geometric model with a cure fraction (G-CF model), we consider the deviance information criterion (DIC) proposed by Spiegelhalter *et al.* (2002) and the expected Akaike information criterion (EAIC) and expected Bayesian information criterion (EBIC) (Brooks, 2002), whose results are shown in Table 4. The better model corresponds to lower DIC, EAIC, and EBIC values.

The criteria above shows that the ZIG-CF model is best. We then select the ZIG-CF model as our working model taking into account the PSBF and criteria in Table 4. For both models, we observe

Table 5: Posterior summaries of the parameters for the fitted ZIG-CF model for the cutaneous melanoma data

Parameter	Estimates		Standard deviation	90% HPD interval
	Mean	Median		
ϕ	0.382	0.381	0.174	(0.099; 0.664)
γ_1	2.056	2.053	0.157	(1.808; 2.317)
γ_2	0.069	0.066	0.029	(0.018; 0.113)
β_0	−1.263	−1.341	0.744	(−2.447; −0.160)
β_2	0.017	0.017	0.009	(0.001; 0.031)
β_3	0.622	0.621	0.117	(0.427; 0.808)

ZIG-CF = zero-inflated geometric model with a cure fraction; HPD = highest posterior density.



(a) ZIP-CF model

(b) ZIG-CF model

Figure 3: *Index plots of ψ-divergence measures for the cutaneous melanoma data. ZIP-CF = zero-inflated Poisson model with a cure fraction; ZIG-CF = zero-inflated geometric model with a cure fraction.*

that, with the exception of covariates age (x_2) and nodule category (x_3), the others had no significant effect. The results for the significant covariates are presented in Table 5, which shows the posterior summary statistics for ZIG-CF model parameters.

In order to detect possible influential observations and verify the model robustness in the presence of these observations, we calculate the Monte Carlo estimates of the measure of divergence ψ for both fitted models. The plots in Figure 3 show the results for the divergence measures KL-, J-, L_1-, and χ^2-distance obtained from the posterior samples of the ZIP-CF and ZIG-CF models parameters. It can be noted, particularly from Figure 3(a), that the observation of index 176 presents a higher value for all divergence measures ψ when compared to other observations that can then be indicated as a possible influential observation. Table 6 condenses the values of divergence measures for the observation of index 176 (referring to a 53-year-old male patient with 4 lymph nodes involved in the disease and who died 6 years after the disease was detected) considering the ZIP-CF and ZIG-CF models. The results show that considering the ZIP-CF model the observation of index 176 is an influential observation since we obtained large values for the divergence measures (larger than the cut-off value). However, the values of the divergence measures for the ZIG-CF model were below the cut-off value and indicated that this model is more robust than the ZIP-CF model in the presence of extreme observations. This conclusion is corroborated by Figure 3(b).

The Bayesian estimate of the deterministic proportion of zero-risk, $\phi/(1 + \phi)$, is 0.276 (sd. =

Table 6: Values of ψ-divergence measures for the observation of index 176 for ZIP-CF and ZIG-CF models

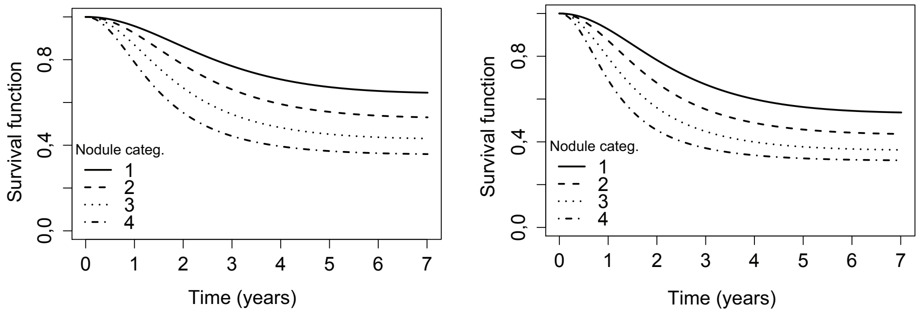
Model	Divergence measures			
	D_{KL}	D_J	D_{L_1}	D_{χ^2}
ZIP-CF	0.262	0.561	0.287	0.608
ZIG-CF	0.161	0.341	0.224	0.341

ZIP-CF = zero-inflated Poisson model with a cure fraction; ZIG-CF = zero-inflated geometric model with a cure fraction.

Table 7: Bayesian criteria for the ZIG-CF and G-CF models.

Criteria	Model	
	ZIG-CF	G-CF
DIC	1023.458	1026.301
EAIC	1030.344	1031.338
EBIC	1054.543	1051.504

ZIG-CF = zero-inflated geometric model with a cure fraction; DIC = deviance information criterion; EAIC = expected Akaike information criterion; EBIC = expected Bayesian information criterion.



(a) 32 years

(b) 66 years

Figure 4: *Bayesian survival function estimate under the ZIG-CF model stratified by nodule category for patients with ages (a)* 32 *and (b)* 66 *years. ZIG-CF = zero-inflated geometric model with a cure fraction.*

0.0987) indicates that 27.6% of the individuals are no longer susceptible (immune) to cutaneous melanoma due to characteristics intrinsic that facilitate the cure of the disease. In addition, the significant value of the estimate of ϕ, $\hat{\phi} = 0.382$, indicates overdispersion in the data. We have also performed a hypothesis test to verify that the frailty variable can be modeled by a geometric distribution instead of a ZIG distribution. In this sense, we may test the hypotheses $H_0 : \phi = 0$ versus $H_1 : \phi \neq 0$. Thus, we estimate the PSBF of the geometric survival model with a cure fraction (M_0 model) with respect to the ZIG-CF model (M_1 model), whose value has resulted in $\mathrm{PSBF}_{01} = 0.312$. This result indicates that the geometric distribution is unsuitable for modeling the frailty variable for cutaneous melanoma data. Table 7 displays the estimates of the Bayesian criteria DIC, EAIC, and EBIC for the ZIG-CF model and geometric model with a cure fraction (G-CF model).

Figure 4 shows the survival function stratified by nodule category (1 to 4) for patients between the ages 32 of 66 years, which correspond to quantiles of 10% and 90% of ages. From these plots, we can observe that the survival probability is seen to decrease more rapidly for older patients and the highest nodule category.

Finally, we analyze the role of the covariates on the proportion of zero-risk (cured fraction) p_0. The positive sign of β_2, coefficient in Table 5, means that the cured fraction decreases with increasing age of patient. Moreover, higher values of nodule category imply smaller cured fraction estimates since

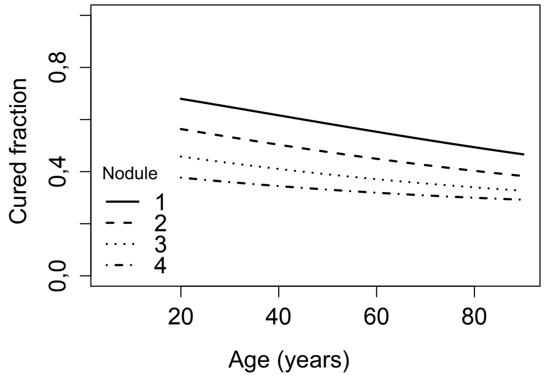


Figure 5: *Cured fraction for the ZIG-CF model versus age stratified by nodule category. ZIG-CF = zero-inflated geometric model with a cure fraction.*

Table 8: Posterior summary of the cured fraction stratified by nodule category and patient age

Age	Nodule category	Estimate		Standard deviation	90% HPD interval
		Mean	Median		
32	1	0.6573	0.6661	0.0732	(0.5429; 0.7787)
	2	0.5549	0.5569	0.0553	(0.4686; 0.6470)
	3	0.4644	0.4650	0.0421	(0.3999; 0.5366)
	4	0.3957	0.3965	0.0459	(0.3169; 0.4656)
66	1	0.5652	0.5667	0.0582	(0.4795; 0.6718)
	2	0.4718	0.4708	0.0432	(0.3991; 0.5414)
	3	0.3995	0.3998	0.0437	(0.3287; 0.4712)
	4	0.3497	0.3515	0.0557	(0.2604; 0.4378)

HPD = highest posterior density.

$\beta_3 > 0$ (Table 5). Figure 5 presents the combined effect of these covariates (age and nodule category) on the cured fraction. Table 5 provides the estimates used to obtained the Bayesian estimates and the 90% HPD interval for the cured fraction (p_0) (Table 8).

7. Conclusions

In this paper, we proposed a new model for accommodating long-term survival data obtained from a discrete frailty model. We have assumed a ZIPS distribution for the frailty variable, that allows accommodating possible overdispersion present in data. Another advantage of our model is that we can classify the proportion of individuals with zero-risk in two components (immune and cured), one is due to inherent characteristics of individuals (deterministic factors) and the other due to random factors, which is not possible in commonly used cure rate models. Moreover, the proposed ZIPS-CF model includes as particular cases some models of literature, such as the models proposed by Chen *et al.* (1999) and Cancho *et al.* (2013).

One important component in this study is to use an appropriate MCMC method. The estimation of the parameters of the proposed model and corresponding inference is conducted by a MCMC algorithm in a Bayesian framework. This is advantageous because we can make statistical inferences based on the posterior quantities of the model, where it is often harder to obtain the standard error of estimated parameters through alternative methods. A widely used algorithm is the Gibbs sampler, which samples iteratively from posterior conditional distributions. However, the Metropolis algorithm should be employed when these conditional distributions are not standard distributions. Thus, we have performed Metropolis-Hasting steps within the Gibbs sampler to obtain samples of the parameters of interest. We also have developed case influence diagnostics for the joint posterior distribution based on the measure of divergence ψ, which has several divergence measures as particular cases, such as L_1-distance, χ^2-distance, J, and KL divergence measures. A model selection has been discussed based on the PSBF, which showed a satisfactory discrimination power between the models. The potentiality of the Bayesian methodology and the applicability of the ZIPS-CF model have been demonstrated with a real cutaneous melanoma dataset, where we have realized that the ZIG-CF model delivers the best fit; in addition, it provides greater robustness in the presence of influential observations. The model proposed in this paper included parametric specification of the baseline distribution; however, the methodology is not restricted to it and some other distributions can be considered.

References

Adamidis K and Loukas S (1998). A lifetime distribution with decreasing failure rate, *Statistics & Probability Letters*, **39**, 35–42.

Ata N and Özel G (2013). Survival functions for the frailty models based on the discrete compound Poisson process, *Journal of Statistical Computation and Simulation*, **83**, 2105–2116.

Barral AM (2001). Immunological studies in malignant melanoma: importance of TNF and the thioredoxin system (Doctorate Thesis), Linkoping University, Linkoping.

Barreto-Souza W, De Morais AL, and Cordeiro GM (2011). The Weibull-geometric distribution, *Journal of Statistical Computation and Simulation*, **81**, 645–657.

Barriga GDC and Louzada F (2014). The zero-inflated Conway-Maxwell-Poisson distribution: Bayesian inference, regression modeling and influence diagnostic, *Statistical Methodology*, **21**, 23–34.

Berkson J and Gage RP (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association*, **47**, 501–515.

Boag JW (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society Series B*, **11**, 15–53.

Brooks SP (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde (2002). *Journal of the Royal Statistical Society Series B*, **64**, 616–618.

Cancho VG, Louzada F, and Ortega EM (2013). The power series cure rate model: an application to a cutaneous melanoma data, *Communications in Statistics-Simulation and Computation*, **42**, 586–602.

Caroni C, Crowder M, and Kimber A (2010). Proportional hazards models with discrete frailty, *Lifetime Data Analysis*, **16**, 374–384.

Casella G and George EI (1992). Explaining the Gibbs sampler, *The American Statistician*, **46**, 167–174.

Chahkandi M and Ganjali M (2009). On some lifetime distributions with decreasing failure rate, *Computational Statistics & Data Analysis*, **53**, 4433–4440.

Chen MH, Ibrahim JG, and Sinha D (1999). A new Bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association*, **94**, 909–919.

Chib S and Greenberg E (1995). Understanding the Metropolis-Hastings algorithm, *The American Statistician*, **49**, 327–335.

Cho H, Ibrahim JG, Sinha D, and Zhu H (2009). Bayesian case influence diagnostics for survival models, *Biometrics*, **65**, 116–124.

Choo-Wosoba H, Levy SM, and Datta S (2015). Marginal regression models for clustered count data based on zero-inflated Conway-Maxwell-Poisson distribution with applications, *Biometrics*, **2**, 606–618.

Consul PC and Jain GC (1973). A generalization of the Poisson distribution, *Technometrics*, **15**, 791–799.

Cook RD and Weisberg S (1982). *Residuals and Influence in Regression*, Chapman and Hall, New York.

Cordeiro GM, Cancho VG, Ortega EMM, and Barriga GDC (2016). A model with long-term survivors: negative binomial Birnbaum-Saunders, *Communications in Statistics-Theory and Methods*, **45**, 1370–1387.

Coskun K (2007). A new lifetime distribution, *Computational Statistics* & *Data Analysis*, **51**, 4497–4509.

Cowles MK and Carlin BP (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association*, **91**, 883–904.

del Castillo J and Pérez-Casany M (2005). Overdispersed and underdispersed Poisson generalizations, *Journal of Statistical Planning and Inference*, **134**, 486–500.

Dey DK and Birmiwal LR (1994). Robust Bayesian analysis using divergence measures. *Statistics* & *Probability Letters*, **20**, 287–294.

Eudes AM, Tomazella VLD, and Calsavara VF (2013). Modelagem de sobrevivência com fração de cura para dados de tempo de vida Weibull modificada, *Revista Brasileira de Biometria*, **30**, 326–342.

Geisser S and Eddy WF (1979). A predictive approach to model selection, *Journal of the American Statistical Association*, **74**, 153–160.

Gelfand AE, Dey DK, and Chang H (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics: Proceedings of the Fourth Valencia International Meeting*, April 15–20, 1991, volume 4, pages 147–167. Oxford University Press, USA.

Gupta PL, Gupta RC, and Tripathi RC (1995). Inflated modified power series distributions with applications, *Communications in Statistics-Theory and Methods*, **24**, 2355–2374.

Hougaard P (1986). A class of multivariate failure time distributions, *Biometrika*, **73**, 671–678.

Ibrahim JG, Chen MH, and Sinha D (2005). *Bayesian Survival Analysis*, Springer, New York.

Kirkwood JM, Ibrahim JG, Sondak VK, *et al.* (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: first analysis of intergroup trial E1690/S9111/C9190, *Journal of Clinical Oncology*, **18**, 2444–2458.

Li CS, Taylor JMG, and Sy JP (2001). Identifiability of cure models, *Statistics* & *Probability Letters*, **54**, 389–395.

Milani EA, Tomazella VLD, Dias TCM, and Louzada F (2015). The generalized time-dependent logistic frailty model: an application to a population-based prospective study of incident cases of lung cancer diagnosed in Northern Ireland, *Brazilian Journal of Probability and Statistics*, **29**, 132–144.

Moger TA, Aalen OO, Halvorsen TO, Storm HH, and Tretli S (2004). Frailty modelling of testicular cancer incidence using Scandinavian data, *Biostatistics*, **5**, 1–14.

Morel JG and Neerchal NK (2012). Overdispersion models in SAS. SAS Institute Inc., Cary, NC.

Morita LHM, Tomazella VL, and Louzada-Neto F (2016). Accelerated lifetime modelling with frailty in a non-homogeneous Poisson process for analysis of recurrent events data, *Quality Technology* & *Quantitative Management*, 1–21.

Ortega EMM, Cordeiro GM, Campelo AK, Kattan MW, and Cancho VG (2015). A power series beta

Weibull regression model for predicting breast carcinoma, *Statistics in Medicine*, **34**, 1366–1388.

Peng F and Dey DK (1995). Bayesian analysis of outlier problems using divergence measures, *Canadian Journal of Statistics*, **23**, 199–213.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues J, Cancho VG, de Castro M, and Louzada-Neto F (2009a). On the unification of long-term survival models, *Statistics & Probability Letters*, **79**, 753–759.

Rodrigues J, de Castro M, Cancho VG, and Balakrishnan N (2009b). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data, *Journal of Statistical Planning and Inference*, **139**, 3605–3611.

Samani EB, Amirian Y, and Ganjali M (2012). Likelihood estimation for longitudinal zero-inflated power series regression models, *Journal of Applied Statistics*, **39**, 1965–1974.

Spiegelhalter DJ, Best NG, Carlin BP, and van der Linde A (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society Series B*, **64**, 583–639.

Sun FB and Kececloglu DB (1999). A new method for obtaining the TTT-plot for a censored sample. In *Proceedings of the Annual Reliability and Maintainability Symposium*, 112–116.

Tahmasbi R and Rezaei S (2008). A two-parameter lifetime distribution with decreasing failure rate, *Computational Statistics & Data Analysis*, **52**, 3889–3901.

Tsodikov AD, Ibrahim JG, and Yakovlev AY (2003). Estimating cure rates from survival data: an alternative to two-component mixture models, *Journal of the American Statistical Association*, **98**, 1063–1078.

Van den Broek J (1995). A score test for zero inflation in a Poisson distribution, *Biometrics*, **51**, 738–743.

Vaupel JW, Manton KG, and Stallard E (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**, 439–454.

Weiss R (1996). An approach to Bayesian sensitivity analysis, *Journal of the Royal Statistical Society Series B*, **58**, 739–750.

Wienke A (2010). *Frailty Models in Survival Analysis*, Chapman and Hall/CRC, New York.

Yakovlev AY and Tsodikov AD (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, New Jersey.

Yang Z, Hardin JW, Addy CL, and Vuong QH (2007). Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model, *Biometrical Journal*, **49**, 565–584.

Received February 20, 2018; Revised August 12, 2018; Accepted August 21, 2018