Brazilian Journal of Probability and Statistics
2013, Vol. 27, No. 4, 585–600
DOI: 10.1214/12-BJPS186

Long-term survival models with latent activation under a flexible family of distributions

Vicente G. Cancho[a], **Mário de Castro**[a] and **Dipak K. Dey**[b]

[a]*Universidade de São Paulo*
[b]*University of Connecticut*

Abstract. In this paper, we propose a new cure rate survival model under a flexible family of distributions. Our approach enables different underlying activation mechanisms that lead to the event of interest. The number of competing causes of the event of interest follows a power series distribution. This model includes the standard mixture cure model and the promotion time cure model. The model is parametrized in terms of the cured fraction, which is then linked to covariates. We carried out a simulation study to assess some properties of our proposal. An illustrative example with a real data set is provided to illustrate the models.

1 Introduction

Models for survival data with a surviving fraction (also known as cure rate models or long-term survival models) have deserved a great deal of interest in the literature under the headings of reliability and survival analysis. Cure rate models cover the situations in that there are sampling units insusceptible to the occurrence of the event of interest. The proportion of such units is termed as the cured fraction. In clinical studies the event of interest may be the death of a patient (which can occur due to different competing causes) or a tumor recurrence (which can be attributed to metastasis-component tumor cells left active after an initial treatment). The literature on the subject is by now vast and expanding rapidly. The books by Maller and Zhou (1996) and Ibrahim et al. (2001), as well as the articles by Chen et al. (1999), Tsodikov et al. (2003), Cooner et al. (2007), Tournoud and Ecochard (2007), de Castro et al. (2009), Rodrigues et al. (2009a), Cancho et al. (2011) and Kim et al. (2011) can be mentioned as few examples.

Two formulations of cure rate models stand out in the literature as being the prevailing approaches. Here we point out a distinguishing feature between them. In the standard mixture cure model (Boag, 1949; Berkson and Gage, 1952), the number of causes of the event of interest is a binary random variable on $\{0, 1\}$, whereas in the promotion time cure model (Yakovlev and Tsodikov, 1996) this number follows a Poisson distribution. These models have been successfully applied to many real word problems. Our aim consists in pursuing some steps toward flexibility.

Key words and phrases. Competing risks, cure rate models, cured fraction, long-term survival models, power series distribution.

Received October 2011; accepted January 2012.

Another approach, introduced by Cooner et al. (2006, 2007), forms an arranged stochastic sequence of latent causes, which induce the occurrence of the event of interest through underlying activation mechanisms. In this paper, we follow this setup and the number of competing causes is modeled by an power series distribution (Johnson et al., 2005). An advantage of our modeling is that the power series distribution is flexible, because it includes as particular cases the Bernoulli, geometric and Poisson distributions, amongst others, which can be tested for the best fitting in a straightforward way.

Since in many applications the cured fraction is of great relevance, in our formulation the models are parametrized in the cured fraction. So, the role of the covariates has the same interpretation, whichever the distribution of the number of competing causes.

As we will see in Section 5, the best fitting is not the one achieved with a Bernoulli or a Poisson distribution for the number of causes of the event of interest. Our example emphasizes the need for flexible models, as allowed by our proposal.

Our paper is organized as follows. In Section 2, we formulate the power series cure rate model. Inference methods based on the likelihood occupy the Section 3. A simulation study with the different models is presented in Section 4. An application to a real data set is developed in Section 5. Finally, Section 6 concludes with some general remarks.

2 Model formulation

For an individual in the population, let M denote the number of causes of the event of interest for this individual. If M is known, we have the competing risks scenario (Klein and Moeschberger, 2003). Here we assume that M is unknown and follows a power series distribution (Johnson et al., 2005) with probability mass function

$$P(M = m; \theta) = \frac{a_m \theta^m}{A(\theta)}, \qquad m = 0, 1, 2, \ldots, \theta > 0, \tag{2.1}$$

where $a_m \geq 0$ and $A(\theta) = \sum_{m=0}^{\infty} a_m \theta^m$. In (2.1) θ and $A(\theta)$ are called the power parameter and the series function, respectively. The probability generating function of M is given by

$$G_M(s) = \frac{A(\theta s)}{A(\theta)} \qquad \text{for } |s| < 1. \tag{2.2}$$

Some distributions of importance belonging to this class are the binomial, Poisson, negative binomial and logarithmic distributions. For example, if k is a positive integer, $a_m = \binom{k}{m}$ and $A(\theta) = (1 + \theta)^k$, then (2.1) defines the binomial distribution with odds equal to θ. A few more examples to be used in the forthcoming sections are as follows:

$$
\begin{aligned}
&a_m = 1/(m+1) \quad \text{and} \quad A(\theta) = -\log(1-\theta)/\theta, 0 < \theta < 1: \qquad \text{logarithmic,}\\
&a_m = \binom{k+m-1}{k-1} \quad \text{and}\\
&A(\theta) = (1-\theta)^{-k}, 0 < \theta < 1: \qquad \text{negative binomial,}\\
&a_m = 1/m! \quad \text{and} \quad A(\theta) = e^{\theta}, \theta > 0: \qquad \text{Poisson,}
\end{aligned}
\tag{2.3}
$$

where k is a positive integer.

Define $\{Z_j\}_{j\in\mathbb{N}}$ being a family of i.i.d. positive continuous random variables independent of a discrete random variable M following a power series distribution, where Z_j is the time corresponding to jth cause to produce the event of interest, with cumulative distribution function $F(z)$ and surviving function $S(z) = 1 - F(z)$. The observable time to event is defined by the random variable $Y = Z_{(R)}$, where R depends on M, $Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(R)} \le \cdots \le Z_{(M)}$ are the order statistics and $Y = \infty$ if $M = 0$. In many biological processes R can be interpreted as a resistance factor of the immune system of the individual. If the event of interest occurs (e.g., cancer relapse), then the random variable Y takes the value of the Rth order statistics $Z_{(R)}$. In other words, as in Cooner et al. (2007), R out of M causes are required to produce the event of interest. The resistance factor can be a fixed constant, a function of M or a random variable specified through a conditional distribution on M. Thus, each different activation mechanism carries a different biological concept.

Using the terminology borrowed from Cooner et al. (2007), in this paper we deal with three specifications for R. First, we assume that given $M \ge 1$, the conditional distribution of R is uniform on $\{1, 2, \dots, M\}$ (random activation scheme). Under this setup, the surviving function for the population is given by

$$
\begin{aligned}
S_{\text{pop}}(y) &= P(Y > y) = P(M = 0) + \{1 - P(M = 0)\}S(y)\\
&= \frac{a_0}{A(\theta)} + \left(1 - \frac{a_0}{A(\theta)}\right)S(y),
\end{aligned}
\tag{2.4}
$$

which comes out to be a mixture cure model with cured fraction $p_0 = P(M = 0) = \lim_{y\to\infty} S_{\text{pop}}(y) = a_0/A(\theta)$. From (2.4), the density function is

$$
f_{\text{pop}}(y) = -S'_{\text{pop}}(y) = \left(1 - \frac{a_0}{A(\theta)}\right)f(y),
$$

where $f(y) = -S'(y)$ denotes the proper density function of the time to event Z. Furthermore, the corresponding hazard function is

$$
h_{\text{pop}}(y) = \frac{(1 - a_0/A(\theta))f(y)}{a_0/A(\theta) + (1 - a_0/A(\theta))S(y)}.
$$

Note that the $f_{\text{pop}}(y)$ and $h_{\text{pop}}(y)$ are improper functions, since $S_{\text{pop}}(y)$ is not a proper surviving function.

As a second setup, the so-called first activation scheme, we suppose that the event of interest happens due to any one of the possible causes. Therefore, for $R = 1$, the time to event is $Y = Z_{(1)} = \min\{Z_1, \dots, Z_M\}$, implying that (Tsodikov et al., 2003; Rodrigues et al., 2009b)

$$S_{\text{pop}}(y) = G_M(S(y)) = \frac{A(\theta S(y))}{A(\theta)}, \tag{2.5}$$

where $G_M(\cdot)$ is as in (2.2). The cured fraction is given by $p_0 = a_0/A(\theta)$. The density function associated to (2.5) is given by

$$f_{\text{pop}}(y) = \frac{A'(\theta S(y))}{A(\theta)} \theta f(y),$$

where $A'(\theta S(y)) = dA(v)/dv|_{v=\theta S(y)}$, with hazard function

$$h_{\text{pop}}(y) = \frac{A'(\theta S(y))}{A(\theta S(y))} \theta f(y).$$

In our third scenario, also known as the last activation scheme, the event of interest only takes place after all the M causes have been occurred, so that $R = M$ and the observed failure time is $Y = Z_{(M)} = \max\{Z_1, \dots, Z_M\}$. According to Cooner et al. (2007),

$$S_{\text{pop}}(y) = 1 + G_M(0) - G_M(F(y)).$$

Hence, from (2.2) we have

$$S_{\text{pop}}(y) = 1 + \frac{a_0}{A(\theta)} - \frac{A(\theta F(y))}{A(\theta)}, \tag{2.6}$$

so that the cured fraction is $p_0 = a_0/A(\theta)$. The surviving function in (2.6) leads to the density function

$$f_{\text{pop}}(y) = \frac{A'(\theta F(y))}{A(\theta)} \theta f(y), \tag{2.7}$$

with hazard function

$$h_{\text{pop}}(y) = \frac{A'(\theta F(y))}{A(\theta) + a_0 - A(\theta F(y))} \theta f(y).$$

The (proper) surviving function for the noncured population, denoted by S_{nc}, is computed by $S_{\text{nc}}(y) = P(Y > y | M \geq 1)$.

From the distribution for the noncured population under the first activation scheme, considering different choices for the distribution of the latent random variables Z_j's, some recently proposed lifetime models can be obtained as special cases. For example, if the $\{Z_j\}_{j \in \mathbb{N}}$ follow the exponential or Weibull distribution, the exponential power series (Chahkandi and Ganjali, 2009) and the Weibull

Table 1 *Non-cured surviving function* (S_{nc}) *and density function* (f_{nc}) *for some models under different activation devices*

Model	Activation	$S_{\text{nc}}(y)$	$f_{\text{nc}}(y)$
Logarithmic	First	$\frac{\log(1-\theta S(y))}{\log(1-\theta)}$	$-\frac{\theta f(y)}{\log(1-\theta)\{1-\theta S(y)\}}$
	Last	$1-\frac{\log(1-\theta F(y))}{\log(1-\theta)}$	$-\frac{\theta f(y)}{\log(1-\theta)\{1-\theta F(y)\}}$
	Random	$S(y)$	$f(y)$
Geometric	First	$\frac{S(y)}{1-\theta F(y)}$	$(1-\theta)f(y)\{1-\theta S(y)\}^{-2}$
	Last	$\frac{(1-\theta)F(y)}{1-\theta F(y)}$	$(1-\theta)f(y)\{1-\theta F(y)\}^{-2}$
	Random	$S(y)$	$f(y)$
Poisson	First	$\frac{e^{-\theta F(y)}-e^{-\theta}}{1-e^{-\theta}}$	$\frac{\theta f(y)e^{-\theta F(y)}}{1-e^{-\theta}}$
	Last	$\frac{1-e^{-\theta S(y)}}{1-e^{-\theta}}$	$\frac{\theta f(y)e^{-\theta S(y)}}{1-e^{-\theta}}$
	Random	$S(y)$	$f(y)$

power series (Morais and Barreto-Souza, 2011) distributions, respectively, are obtained. Rodrigues et al. (2011) formulate a flexible density function from a selection mechanism viewpoint. Within this approach, we find for example the exponential power series distribution as a special case. Also, from the distribution for the noncured population under the last activation scheme, if the distribution of the variables Z_j's is exponential, then we get the complementary exponential power series distribution proposed by Flores et al. (2013). This family of distributions includes as particular cases the distributions proposed by Cancho et al. (2011) and Louzada-Neto et al. (2011).

We present in Table 1 the density and surviving functions for the noncured population under different activation schemes. From Table 1 new families of distributions can be generated. Under the first activation scheme, if M follows a geometric distribution and the variables Z_j's follow an exponential or a Weibull distribution, we obtain the distributions introduced by Marshall and Olkin (1997) in Sections 3 and 4, respectively.

In many applications of long-term survival models the cured fraction plays a central role. With this concern in mind, we change the parametrization of the model in order to put the cured fraction p_0 in the expressions. Since $p_0 = P(M = 0) = a_0/A(\theta)$, we have $\theta = A^{-1}(a_0/p_0)$. In the sequel, we will work with the logarithmic, geometric and Poisson distributions. For these distributions, remembering (2.3), we obtain $a_0 = 1$ and computing $A^{-1}(\theta)$ we get $1 + W(-\theta e^{-\theta})/\theta$, $1 - 1/\theta$ and $\log(\theta)$, respectively, where $W(\cdot)$ stands for the Lambert W function (Corless et al., 1996). As we will see more clearly in Section 3, this parametrization is advantageous. Using p_0 as parameter and the expressions (2.4)–(2.7), we arrive at the improper surviving and density functions presented in Table 2. From these results, we realize that, whichever the activation scheme, the cured fraction is the

Table 2 *Surviving function* (S_{pop}) *and density function* (f_{pop}) *for some models under different activation devices*

Model	Activation	$S_{\text{pop}}(y)$	$f_{\text{pop}}(y)$
Logarithmic	First	$-\frac{\log(1-W_0S(y))}{W_0S(y)}p_0$	$\frac{W_0S(y)+\{1-W_0S(y)\}\log(1-W_0S(y))}{\{1-W_0S(y)\}W_0S(y)^2}p_0f(y)$
	Last	$1+p_0+\frac{\log(1-W_0F(y))}{W_0F(y)}p_0$	$\frac{W_0F(y)+\{1-W_0F(y)\}\log(1-W_0F(y))}{\{1-W_0F(y)\}W_0F(y)^2}p_0f(y)$
	Random	$p_0+(1-p_0)S(y)$	$(1-p_0)f(y)$
Geometric	First	$\{1+(p_0^{-1}-1)F(y)\}^{-1}$	$\frac{p_0^{-1}-1}{\{1+(p_0^{-1}-1)F(y)\}^2}f(y)$
	Last	$1+p_0-\{1+(p_0^{-1}-1)S(y)\}^{-1}$	$\frac{p_0^{-1}-1}{\{1+(p_0^{-1}-1)S(y)\}^2}f(y)$
	Random	$p_0+(1-p_0)S(y)$	$(1-p_0)f(y)$
Poisson	First	$p_0^{F(y)}$	$-\log(p_0)p_0^{F(y)}f(y)$
	Last	$1+p_0-p_0^{S(y)}$	$-\log(p_0)p_0^{S(y)}f(y)$
	Random	$p_0+(1-p_0)S(y)$	$(1-p_0)f(y)$

Remark: $W_0 = 1 + p_0W(-e^{-1/p_0}/p_0)$, where $W(\cdot)$ is the Lambert W function (Corless et al., 1996).

same. The models differ by its surviving, density and hazard functions. Moreover, according to the Theorem 2.1 in Kim et al. (2011), we have that $S_{\text{pop}}(y)$ in (2.4) $\geq S_{\text{pop}}(y)$ in (2.5). In a similar way, it can be shown that $S_{\text{pop}}(y)$ in (2.6) $\geq S_{\text{pop}}(y)$ in (2.4).

Hereafter, we assume a Weibull distribution for the unobserved time Z with $F(z;\boldsymbol{\gamma}) = 1-\exp(-z^{\gamma_1}e^{\gamma_2})$ and $f(z;\boldsymbol{\gamma}) = \gamma_1 z^{\gamma_1-1}\exp(\gamma_2 - z^{\gamma_1}e^{\gamma_2})$, for $z > 0$, $\gamma_1 > 0$, $\gamma_2 \in \mathbb{R}$ and $\boldsymbol{\gamma} = (\gamma_1,\gamma_2)^\top$. Figure 1 portrays distinct behaviors of the surviving functions in Table 2. These plots illustrate the flexibility afforded by our proposal.

3 Inference

Let us consider the situation when the failure time Y in Section 2 is not completely observed and is subject to right censoring. Let C_i denote the censoring time. In a sample of size n, we then observe $T_i = \min\{Y_i, C_i\}$ and $\delta_i = \mathrm{I}(Y_i \leq C_i)$, where $\delta_i = 1$ if T_i is a failure time and $\delta_i = 0$ if it is right censored, for $i = 1,\ldots,n$.

Let $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^\top$ denote the vector of covariates for the ith individual. Completing our model, we propose to relate the cured fraction to the covariates by the logistic link

$$\log\left(\frac{p_{0i}}{1-p_{0i}}\right) = \mathbf{x}_i^\top\boldsymbol{\beta} \quad \text{or} \quad p_{0i} = \frac{\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top\boldsymbol{\beta})}, \tag{3.1}$$

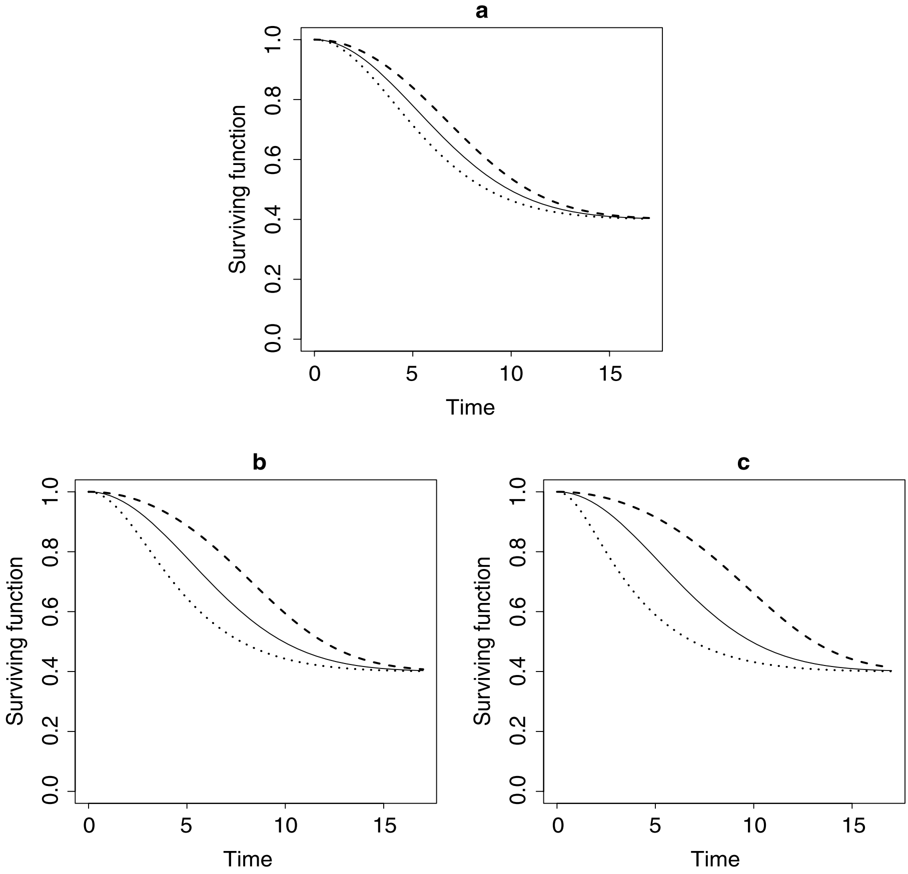

Figure 1 *Surviving functions for the* (a) *Poisson*, (b) *geometric and* (c) *logarithmic models with* $p_0 = 0.4$ *and a Weibull distribution* ($\gamma_1 = 2$, $\gamma_2 = -4$) *under different activations* (*last*: *dashed*, *random*: *solid and first*: *dotted*).

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ encapsulates the vector of regression coefficients, so that for each group of individuals represented by $\mathbf{x}_i$, we have a different cured fraction. With this link function, the models are identifiable in the sense of Li et al. (2001).

We stress that regardless the specific model in Table 2, covariates are associated to the cured fraction through a unique expression very familiar to practitioners. If we had adopted the parametrization in θ, for the Poisson and logarithmic distributions the cured fraction is $p_0 = e^{-\theta}$ and $p_0 = -\theta/\log(1-\theta)$, respectively, and θ would be linked to the covariates (e.g., log and logistic links). The connections between the cured fraction and the covariates would be much more clumsier in these expressions than in (3.1). Therefore, although the entries for the logarithmic model in Table 2 seem uneasy, we have a direct interpretation of the coefficients in (3.1).

With the expression (3.1), we can write the likelihood of $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ under noninformative censoring as

$$L(\boldsymbol{\vartheta}; \mathbf{D}) \propto \prod_{i=1}^{n} f_{\text{pop}}(t_i; \boldsymbol{\vartheta})^{\delta_i} S_{\text{pop}}(t_i; \boldsymbol{\vartheta})^{1-\delta_i}, \qquad (3.2)$$

where $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta}, \mathbf{x})$, $\mathbf{t} = (t_1, \dots, t_n)^\top$, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^\top$, whereas $f_{\text{pop}}(\cdot; \boldsymbol{\vartheta})$ and $S_{\text{pop}}(\cdot; \boldsymbol{\vartheta})$ are the improper density and surviving functions in Table 2.

From the likelihood function in (3.2), the maximum likelihood estimation of the parameter $\boldsymbol{\vartheta}$ is carried out. Numerical maximization of the log-likelihood function $\ell(\boldsymbol{\vartheta}; \mathbf{t}, \boldsymbol{\delta}) = \log(L(\boldsymbol{\vartheta}; \mathbf{t}, \boldsymbol{\delta}))$ is accomplished by using the R language (R Development Core Team, 2011). The Lambert W function in Table 2 can be found in the R package emdbook. The computational program is available from the authors upon request. Under suitable regularity conditions, it can be shown that the asymptotic distribution of the maximum likelihood estimator $\widehat{\boldsymbol{\vartheta}}$ is multivariate normal with mean vector $\boldsymbol{\vartheta}$ and covariance matrix $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\vartheta}})$, which can be estimated by

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\vartheta}}) = \left\{ -\frac{\partial^2 \ell(\boldsymbol{\vartheta}; \mathbf{t}, \boldsymbol{\delta})}{\partial \boldsymbol{\vartheta}\, \partial \boldsymbol{\vartheta}^\top} \right\}^{-1},$$

evaluated at $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$. The first and second derivatives of the log-likelihood function are computed numerically.

Different models can be compared by penalizing over-fitting by using the Akaike information criterion (*AIC*) and the Schwartz Bayesian criterion (*SBC*) given by $AIC = -2\ell(\widehat{\boldsymbol{\vartheta}}) + 2\#(\boldsymbol{\vartheta})$ and $SBC = -2\ell(\widehat{\boldsymbol{\vartheta}}) + \#(\boldsymbol{\vartheta}) \log(n)$, where $\#(\boldsymbol{\vartheta})$ is the number of model parameters. The model with the smallest value of any of these criteria (among all candidate models) is commonly taken as the preferred model for describing the given data set.

4 Simulation study

To evaluate the performance of the estimators of the cured fraction under different models and activation schemes, we carried out a simulation study. In this study, we considered the logarithmic cure rate model under both the first and the last activation schemes as given in Table 2 with the Weibull distribution for the event times (Z) with parameter $\gamma_1 = 2$ and $\gamma_2 = -3.2$. For each individual i, $i = 1, \dots, n$, the number of causes of the event of interest for this individual (M_i) is generated from the logarithmic distribution with parameter θ_i equal to $p_{0i} W(-e^{-1/p_{0i}}/p_{0i})$, where $W(\cdot)$ is the Lambert W function (Corless et al., 1996). In our simulations, we have one binary covariate x with values drawn from a Bernoulli distribution with parameter 0.5. In this way, $p_{0i} = \exp(\beta_1 + \beta_2 x_i)/\{1 + \exp(\beta_1 + \beta_2 x_i)\}$. We took $\beta_1 = -0.5$ and $\beta_2 = 0.7$, so that the cured fraction for the two levels of x are

$p_0^{(0)} = 0.378$ and $p_0^{(1)} = 0.550$, respectively. The censoring times were sampled from the uniform distribution on the interval $(0, \tau)$, where τ was set in order to control the proportion of censored observations. In this study, the proportion of censored observations was on average approximately equal to 57%.

We chose three sample sizes, $n = 200$, 400 and 800. For each configuration, we conducted 1000 simulations and then calculated the average of the maximum likelihood estimatives (MLEs) of the cured fraction ($p_0^{(0)}$ and $p_0^{(1)}$), as well as the standard deviation (SD) of the MLEs and the square root of the mean squared error (RMSE) of the MLEs.

The simulation results are shown in Tables 3 and 4. We can observe that the averages of the maximum likelihood estimates of the cured fraction are close to the true values when the fitted model and the activation scheme are the correct ones. When the model or the activation scheme is incorrect, the estimators of the cured fraction are more biased. However, the bias is more sensitive to the activation scheme than to the distribution of the number of causes. Similar results were ob-

Table 3 *Maximum likelihood estimates average* (*AMLE*), *standard deviation* (*SD*) *and square root of the mean squared error* (*RMSE*) *of the cured fractions* $p_0^{(0)}$ (*left*, *true value* $= 0.378$) *and* $p_0^{(1)}$ (*right*, *true value* $= 0.550$) *for simulated data from the logarithmic cure rate model under the first activation scheme*

Fitted model	Activation	n	AMLE		SD		RMSE	
Logarithmic	First	200	0.380	0.550	0.0489	0.0560	0.0489	0.0560
		400	0.377	0.549	0.0346	0.0400	0.0346	0.0400
		800	0.377	0.549	0.0237	0.0287	0.0237	0.0287
	Last	200	0.451	0.514	0.0452	0.0715	0.0865	0.0798
		400	0.448	0.515	0.0299	0.0502	0.0770	0.0609
		800	0.447	0.511	0.0205	0.0368	0.0720	0.0536
Geometric	First	200	0.368	0.559	0.0550	0.0561	0.0558	0.0568
		400	0.364	0.559	0.0386	0.0398	0.0410	0.0408
		800	0.366	0.559	0.0262	0.0287	0.0286	0.0303
	Last	200	0.425	0.534	0.0512	0.0671	0.0696	0.0689
		400	0.421	0.535	0.0341	0.0471	0.0549	0.0495
		800	0.420	0.532	0.0234	0.0349	0.0484	0.0392
Poisson	First	200	0.366	0.564	0.0591	0.0571	0.0601	0.0589
		400	0.361	0.564	0.0409	0.0403	0.0442	0.0428
		800	0.364	0.565	0.0276	0.0294	0.0309	0.0329
	Last	200	0.397	0.552	0.0569	0.0633	0.0601	0.0633
		400	0.392	0.552	0.0382	0.0443	0.0408	0.0443
		800	0.393	0.551	0.0260	0.0327	0.0303	0.0327
Mixture cure		200	0.374	0.563	0.0606	0.0594	0.0606	0.0608
		400	0.369	0.563	0.0413	0.0417	0.0422	0.0437
		800	0.372	0.563	0.0279	0.0306	0.0285	0.0333

Table 4 *Maximum likelihood estimates average* (*AMLE*), *standard deviation* (*SD*) *and square root of the mean squared error* (*RMSE*) *of the cured fractions* $p_0^{(0)}$ (*left*, *true value* $= 0.378$) *and* $p_0^{(1)}$ (*right*, *true value* $= 0.550$) *for simulated data from the logarithmic cure rate model under the last activation scheme*

Fitted model	Activation	*n*	AMLE		SD		RMSE	
Logarithmic	First	200	0.433	0.460	0.0680	0.0815	0.0876	0.1214
		400	0.437	0.465	0.0425	0.0550	0.0732	0.1007
		800	0.443	0.473	0.0305	0.0393	0.0718	0.0862
	Last	200	0.379	0.547	0.0692	0.0710	0.0692	0.0711
		400	0.377	0.550	0.0453	0.0498	0.0453	0.0498
		800	0.378	0.551	0.0330	0.0348	0.0330	0.0348
Geometric	First	200	0.410	0.465	0.0859	0.0938	0.0919	0.1265
		400	0.418	0.476	0.0483	0.0562	0.0632	0.0929
		800	0.424	0.483	0.0342	0.0397	0.0573	0.0774
	Last	200	0.367	0.545	0.0788	0.0713	0.0794	0.0715
		400	0.367	0.548	0.0521	0.0502	0.0532	0.0502
		800	0.367	0.550	0.0382	0.0352	0.0396	0.0352
Poisson	First	200	0.403	0.493	0.0767	0.0790	0.0808	0.0975
		400	0.406	0.498	0.0483	0.0528	0.0559	0.0737
		800	0.409	0.504	0.0353	0.0379	0.0470	0.0598
	Last	200	0.368	0.535	0.0882	0.0708	0.0887	0.0723
		400	0.367	0.548	0.0521	0.0497	0.0532	0.0497
		800	0.371	0.541	0.0404	0.0355	0.0410	0.0367
Mixture cure		200	0.389	0.518	0.0780	0.0729	0.0788	0.0797
		400	0.389	0.521	0.0509	0.0508	0.0522	0.0583
		800	0.390	0.525	0.0377	0.0366	0.0398	0.0444

served when the number of causes was generated from the Bernoulli and Poisson distributions. For the sake of space, we omit here the tables presented in a supplemental file (Cancho et al., 2012). Furthermore, in most of the situations in these tables, when the fitted model and the activation scheme are the correct ones, there is also a gain in terms of the RMSEs. Once again, an incorrect activation scheme has greater impact than an incorrect distribution. As expected, the SDs and RMSEs decrease as the sample size increases.

5 Application

In this section, we work out an example employing the models presented in Section 2. The data set includes 205 patients observed after operation for removal of malignant melanoma in the period 1962–1977. The patients were followed until 1977. These data are available in the timereg package in R (Scheike, 2009). The observed time (T) ranges from 10 to 5565 days (from 0.0274 to 15.25 years) and refers to the time until the patient's death or the censoring time. Patients dead from other causes, as well as patients still alive at the end of the study are censored observations (72%). We take ulceration status (absent, $n = 115$; present, $n = 90$) and tumor thickness (in mm, mean $= 2.92$ and standard deviation $= 2.96$) as covariates with coefficients β_{ulc} and β_{thick}, respectively, whereas β_1 denotes the intercept. The Kaplan–Meier estimate of the surviving function given in Figure 2 levels off above 0.6. The presence of a plateau indicates that models that ignore the possibility of cure will not be suitable for these data.

Once the distribution of the number of possible causes and the activation scheme are unknown, statistical model fitting could give some guidelines on what they could be. First, we fitted the models described in Table 2. In Table 5, we applied the selection criteria on the candidate models. According to the *AIC* and *SBC*

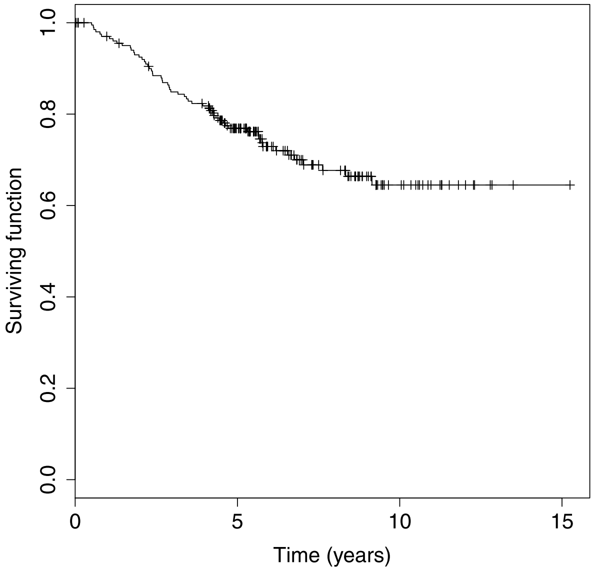

Figure 2 *Kaplan–Meier estimate of the surviving function.*

Table 5 *AIC and SBC for the fitted models*

	Activation scheme					
	First		Last		Random	
Model	*AIC*	*SBC*	*AIC*	*SBC*	*AIC*	*SBC*
Poisson	425.1	441.7	438.5	454.7	431.1	447.7
Geometric	420.8	437.4	492.7	509.3	431.1	447.7
Logarithmic	416.1	433.1	447.2	463.8	431.1	447.7

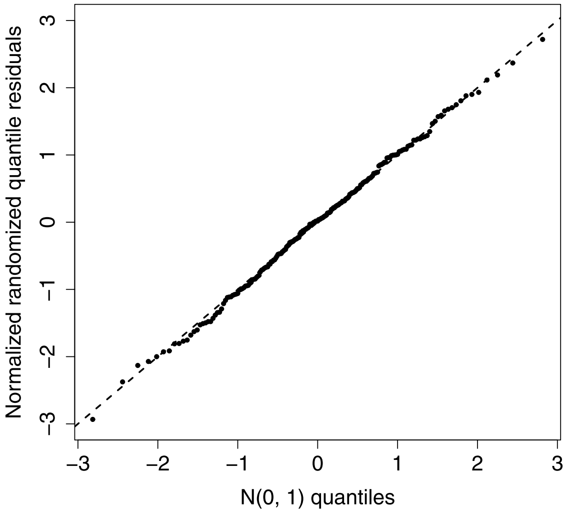

Figure 3 *QQ plot of the normalized randomized quantile residuals with identity line for the Log-*1*st model.*

criteria, the logarithmic cure rate model under the first activation scheme (Log-1st, say) stands out as the best one followed by the geometric model under the first activation scheme (Geo-1st, say). We also tried different negative binomial models in (2.3) by taking $k = 2, 3, \ldots, 10$, but none of these models yielded a better fit than the geometric model ($k = 1$). The QQ plot of the normalized randomized quantile residuals (Dunn and Smyth, 1996; Rigby and Stasinopoulos, 2005) in Figure 3 suggests that the Log-1st model yields an acceptable fit. Each point in Figure 3 corresponds to the median of five sets of ordered residuals. Taking into account the criteria in Table 5 and the QQ plot in Figure 3, we select the Log-1st model as our working model.

Maximum likelihood estimates (MLE) of the coefficients are in Table 4. The estimate of the shape parameter (γ_1) furnishes an evidence against the exponential distribution ($\gamma_1 = 1$) for the unobserved failure times. The covariates have a significant effect on the reduction of the cured fraction. Comparing patients with ulceration absent and present, the odds ratio of the cured fraction is estimated as $e^{1.482} = 4.4$. This estimate is easily computed from Table 6 due to the parametrization in the cured fraction.

Figure 4 displays the surviving function stratified by ulceration status for patients with tumor thickness equal to 0.64, 1.94 and 6.63 mm, which correspond to the 10, 50 and 90 percentiles. These plots highlight the combined impact of the covariates on the cured fraction. Finally, Table 7 brings the MLE of the cured fraction for the plots in Figure 4, as well as the estimates from the Geo-1st model, which is the second best model according to the criteria in Table 5. Approximate 95% con-

Table 6 *Maximum likelihood estimates of the parameters for the Log-1st model*

Parameter	Estimate (est)	Standard error (se)	|est|/se
γ_1	2.211	0.2724	–
γ_2	−4.187	0.5153	8.13
β_1	1.677	0.3564	4.71
β_{ulc}	−1.482	0.3278	4.52
β_{thick}	−0.141	0.0356	3.97

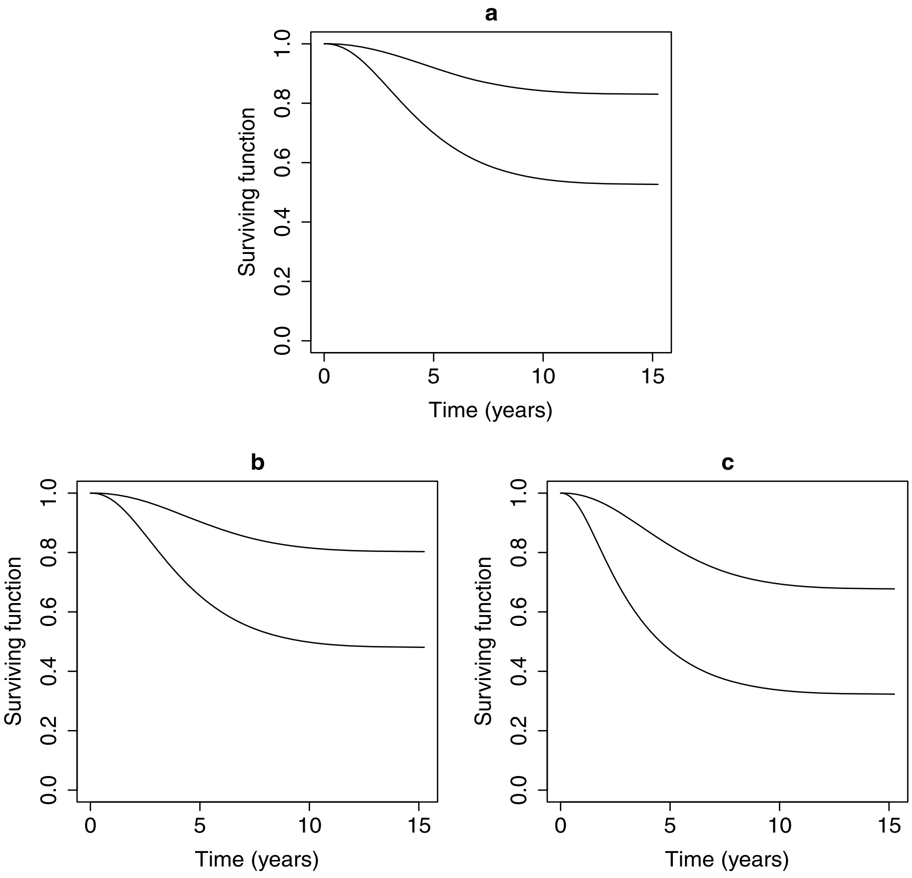

Figure 4 *Surviving function under the Log-1st model stratified by ulceration status (upper: absent, lower: present) for patients with tumor thickness equal to* (a) 0.64, (b) 1.94 *and* (c) 6.63 *mm.*

fidence intervals were obtained after an application of the delta method. There are some differences in the estimates from these models, but for each selected value of tumor thickness the intervals do not overlap.

Table 7 *Maximum likelihood estimates of the cured fraction stratified by ulceration status and selected tumor thickness under the Log-*1*st* (*above*) *and Geo-*1*st* (*below*) *models*

Tumor thickness	Ulceration	Estimate	Standard error	95% confidence interval
0.64	Absent	0.830	0.0489	(0.734, 0.926)
		0.845	0.0442	(0.759, 0.932)
	Present	0.526	0.0794	(0.371, 0.682)
		0.554	0.0844	(0.389, 0.720)
1.94	Absent	0.803	0.0527	(0.700, 0.906)
		0.813	0.0496	(0.716, 0.910)
	Present	0.481	0.0726	(0.338, 0.623)
		0.496	0.0779	(0.344, 0.649)
6.63	Absent	0.677	0.0725	(0.535, 0.819)
		0.652	0.0890	(0.478, 0.827)
	Present	0.306	0.0552	(0.218, 0.428)
		0.299	0.0686	(0.165, 0.433)

6 Conclusions

In this paper, we proposed the power series cure rate model under different activations as a flexible model for modeling survival data with a cured fraction. The model is flexible and includes as particular cases the geometric, Poisson and logarithmic distributions. Moreover, under the random activation scheme the mixture cure model is recovered. The models can be tested for the best fitting in a straightforwardly way. In the application to a melanoma data set, we discovered that the logarithmic cure rate model under the first activation scheme delivers the best fit. We observed that the surviving probability decreases more rapidly for patients with thicker tumors, and that the cured fraction is lower for patients with ulceration. The interpretation of the role of covariates is easy due to the parametrization in the cured fraction.

Our formulation in Section 2 is based on the Weibull distribution for the unobserved time. In the same lines of the works by Kuk and Chen (1992) and Sy and Taylor (2000), for example, we might envision an extension of the present paper to a semiparametric setting.

Acknowledgments

The authors thank a referee for comments which contributed to improve the presentation of the paper. V. G. Cancho supported by CNPq, Brazil. M. de Castro supported by FAPESP, Brazil.

Supplementary Material

Supplement: Long-term survival models with latent activation under a flexible family of distributions (DOI: 10.1214/12-BJPS186SUPP; .pdf). Additional results from the simulation study.

References

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

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

Cancho, V. G., de Castro, M. and Dey, D. K. (2012). Supplement to "Long-term survival models with latent activation under a flexible family of distributions." DOI:10.1214/12-BJPS186SUPP.

Cancho, V. G., Louzada-Neto, F. and Barriga, G. D. C. (2011). The Poisson-exponential lifetime distribution. *Computational Statistics & Data Analysis* **55**, 677–686. MR2736587

Chahkandi, M. and Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate. *Computational Statistics & Data Analysis* **53**, 4433–4440. MR2744336

Chen, M.-H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* **94**, 909–919. MR1723307

Cooner, F., Banerjee, S. and McBean, A. M. (2006). Modelling geographically referenced survival data with a cure fraction. *Statistical Methods in Medical Research* **15**, 307–324. MR2242244

Cooner, F., Banerjee, S., Carlin, B. P. and Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association* **102**, 560–572. MR2370853

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J. and Knuth, D. E. (1996). On the Lambert *W* function. *Advances in Computational Mathematics* **5**, 329–359. MR1414285

de Castro, M., Cancho, V. G. and Rodrigues, J. (2009). A Bayesian long-term survival model parametrized in the cured fraction. *Biometrical Journal* **51**, 443–455. MR2750045

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**, 236–244.

Flores D., J., Borges, P., Cancho, V. G. and Louzada, F. (2013). The complementary exponential power series distribution. *Brazilian Journal of Probability and Statistics* **27**, 565–584.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer. MR1876598

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distributions*, 3rd ed. Hoboken: Wiley. MR2163227

Kim, S., Chen, M.-H. and Dey, D. K. (2011). A new threshold regression model for survival data with a cure fraction. *Lifetime Data Analysis* **17**, 101–122. MR2764586

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis*, 2nd ed. New York: Springer.

Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 531–541.

Li, C. S., Taylor, J. M. and Sy, J. P. (2001). Identifiability of cure models. *Statistics and Probability Letters* **54**, 389–395. MR1861384

Louzada-Neto, F., Roman, M. and Cancho, V. G. (2011). The complementary exponential geometric distribution: Model, properties and a comparison with its counterpart. *Computational Statistics & Data Analysis* **55**, 2516–2524. MR2787009

Maller, R. A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. New York: Wiley. MR1453117

Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika* **84**, 641–652. MR1603936

Morais, A. L. and Barreto-Souza, W. (2011). A compound class of Weibull and power series distributions. *Computational Statistics & Data Analysis* **55**, 1410–1425. MR2741424

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics* **54**, 507–554. MR2137253

Rodrigues, J., de Castro, M., Cancho, V. G. and Balakrishnan, N. (2009a). COM–Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference* **139**, 3605–3611. MR2549108

Rodrigues, J., Cancho, V. G., de Castro, M. and Louzada-Neto, F. (2009b). On the unification of the long-term survival models. *Statistics & Probability Letters* **79**, 753–759. MR2662300

Rodrigues, J., Balakrishnan, N., Cordeiro, G. M. and de Castro, M. (2011). A unified view on lifetime distributions arising from selection mechanisms. *Computational Statistics & Data Analysis* **55**, 3311–3319. MR2825413

Scheike, T. (2009). timereg package. R package version 1.1-0. With contributions from T. Martinussen and J. Silver. R package version 1.1-6.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56**, 227–236. MR1767631

Tournoud, M. and Ecochard, R. (2007). Application of the promotion time cure model with time-changing exposure to the study of HIV/AIDS and other infectious diseases. *Statistics in Medicine* **26**, 1008–1021. MR2339230

Tsodikov, A. D., Ibrahim, J. G. and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association* **98**, 1063–1078. MR2055496

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Singapore: World Scientific.

V. G. Cancho
M. de Castro
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos-SP
Brasil
E-mail: garibay@icmc.usp.br; mcastro@icmc.usp.br

D. K. Dey
Department of Statistics
University of Connecticut
Storrs, Connecticut 06269-4120
USA
E-mail: dipak.dey@uconn.edu