Statistics and Its Interface Volume 12 (2019) 561–571

Destructive power series long-term survival model with complex activation schemes

Diego I. Gallardo[*,†], Heleno Bolfarine,
Antonio C. Pedroso-de-Lima, and Jose S. Romeo

A new destructive cure rate model is introduced based on a family of power series distribution for the number of concurrent causes related to the event of interest. A mixture of first and last activation schemes is considered. For parameter estimation a classical approach based on maximum likelihood methodology is implemented. The performance of estimation procedure is evaluated based on a small scale simulation study. The model is also considered on a real data example, involving congestive heart failure patients.

AMS 2000 subject classifications: Primary 62N01, 62N01; secondary 62N02.
Keywords and phrases: Cure rate models, Competing risks, Power series distribution.

1. INTRODUCTION

Cure rate models have become the *ad-hoc* choice when the event of interest may not be attainable for a fraction of individuals in the population. A possible way to deal with this situation is to consider that there are a random number M of possible concurrent causes of failure, with corresponding latent times given by continuous non-negative random variables $W_1, \dots, W_M$. Conditionally on $M = m$, these quantities are assumed to be independent and identically distributed (*iid*) so that the failure time T is given by

$$T = \begin{cases} \min(W_1, \dots, W_M), & \text{if } M > 0; \\ \infty, & \text{if } M = 0. \end{cases}$$

Different distributions for M and W_j have been extensively considered by several authors. The seminal work by Berkson and Gage [1] assumed the combination Bernoulli/Exponential models. Several decades later, Chen *et al.* [2] considered a Poisson/Weibull structure for the problem. Based on the same Weibull distribution for the latent times, Rodrigues *et al.* [3], [4] proposed a more flexible framework assuming, respectively, Negative Binomial and COM-Poisson distributions for M. Cancho *et al.* [5] considered Geometric/Birnbaum-Saunders models and, later on, Cancho *et al.* [6] studied the combination Power series/Weibull. Negative Binomial/Generalized Gamma and Power series/Beta-Weibull models were considered by Ortega *et al.* [7], [8]. Cordeiro *et al.* [9] examined the Negative Binomial/Birnbaum-Saunders combination and recently Gallardo *et al.* [10] developed the model based on the Yule-Simon/Weibull distributions.

Rodrigues *et al.* [11] elaborated a more general model. Assuming the availability of some intervention, they considered that out of M original risk factors, only a number $D(\leq M)$ remains in effect. For instance, in oncological studies, M usually represents the number of carcinogenic cells for a patient that has some evidence of cancer. After an initial treatment, D of such cells would remain active. Therefore, considering the cure as $M = 0$ (i.e. the patient would not have any remaining carcinogenic cells) would be contradictory. In such a case, cure will be achieved when $D = 0$. In their initial proposal, Rodrigues *et al.* [11] imposed the weighted Poisson distribution for M, with Poisson and Negative Binomial distributions as special cases. Conditionally on $M = m$, the random variable D is assumed to have a Binomial distribution with size m and success probability p, i.e., each initial concurrent causes can be independently activated with probability p. Under the constraint $D \leq M$, one has the *destructive* structure. Other possibilities have been considered elsewhere (see, e.g., Yang and Chen [12]).

Let $W_1, \dots, W_D$ be activation times related to non-destroyed causes, assumed to be conditionally independent (given $D = d$) and identically distributed. The corresponding failure time is then given by $T = \min(W_1, \dots, W_D)$ for $D > 0$ and $T = \infty$ for $D = 0$. This representation is known in the literature as *first activation scheme* (FA). In Cooner *et al.* [13] one can find a more general activation scheme. Specifically, they assume that it is necessary to have activation of a random number of underlying causes, say R, to have the event of interest. Under their definition, $R = 1$ would correspond to the FA scheme. It is conceivable then to consider situations where, instead of the minimum, it would be required to have the maximum among all concurrent times ($R = M$ or $R = D$ in the non-destructive and destructive models, respectively). This scheme is known as the *last activation scheme* (LA). In addition, Cooner *et al.*

[*] Corresponding author.

[†] The first author has been partially supported by FONDECYT (Fondo Nacional de Desarrollo Científico y Tecnológico) Grant 11160670.

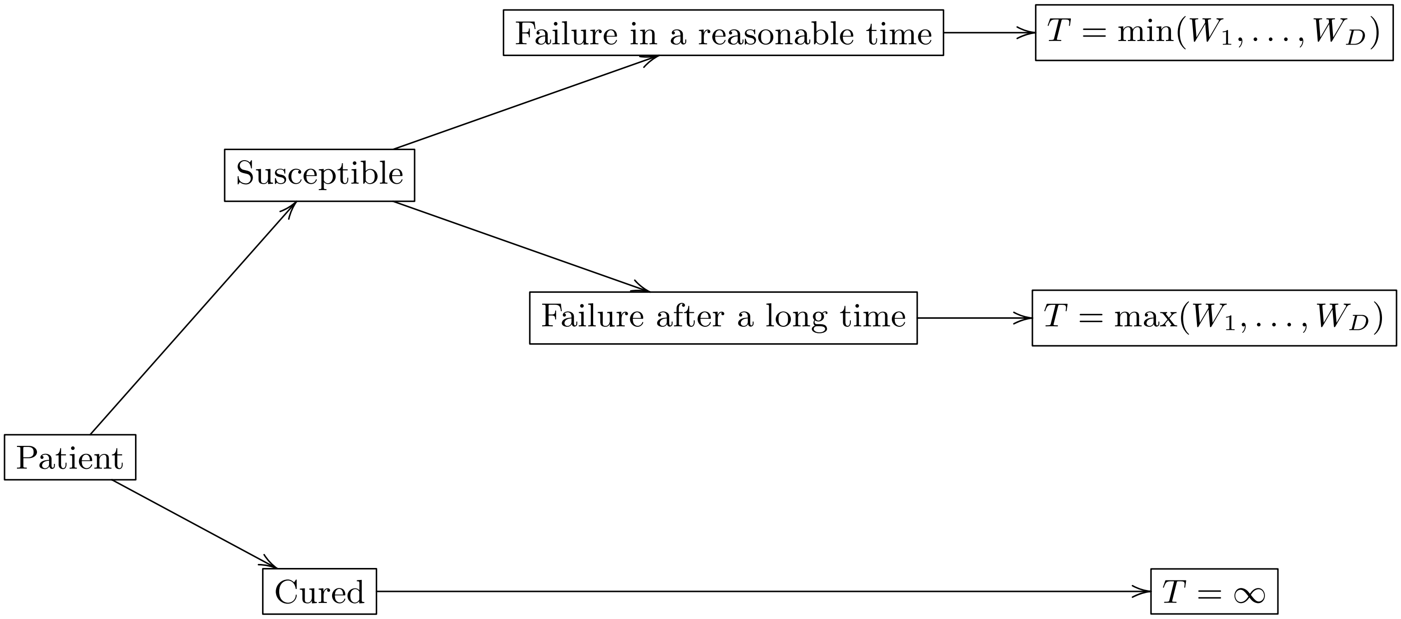
Failure in a reasonable time
$T = \min(W_1, \dots, W_D)$
Susceptible
Failure after a long time
$T = \max(W_1, \dots, W_D)$
Patient
Cured
$T = \infty$

Figure 1. Activation scheme for cured and susceptible individuals.

[13] also propose the so-called *random activation scheme* (RA), assuming discrete uniform distribution for R, similar to the mixture model of Berkson and Gage [1]. The authors also consider a *mixture activation scheme* (mix) which considers that, for each individual, the FA and LA schemes appear with probabilities π and $1 - \pi$, respectively. In the non-destructive context, Cancho *et al.* [14] deal with three of the above mentioned activation schemes (FA, LA and RA) and the power series distribution for the concurrent causes whereas Cancho *et al.* [15] use the same idea in the destructive framework, which was extended in Gallardo *et al.* [16] in a random effects model context. From our knowledge, the different activation schemes considered by Cooner *et al.* [13] have not been applied to destructive models so far. Moreover, our motivation to introduce different activation schemes in a destructive cure rate model context is given by a real data set application related to congestive heart failure (CHF) patients. Specifically, we are interested in introduce the mixture activation scheme in this context. The physicians dealing with subjects under this condition expect survivals no longer than a certain time, say 5 years. However, they have identified patients with an unusually long term follow-up, some of them with lifetimes similar to the general (non-CHF) individuals. It seems reasonable, then, to consider the population of CHF under a 3-fold stratification: patients with lifetime under the 5 year period (FA scheme), subjects with a long-term survival though inferior to the non-CHF individuals (LA scheme) and patients with lifetime following a pattern similar to the general population (“cured” subjects). Figure 1 describes a diagram with this situation. Also, given the possibility of a long follow-up and the active research in the development of new drugs, it is reasonable to consider that the initial number M of activation times may be reduced, as is the case in the destructive model considered in this paper.

Since FA and LA schemes are particular cases for $\pi = 1$ and $\pi = 0$, respectively, we may implement statistical based procedures to decide between them. We propose a new class of destructive cure rate models based on the power series distribution, considering FA, LA and a mixture between those two schemes.

The remaining of this paper is organized as follows. In Section 2, we formulate the new model and special cases are presented. In Section 3, we discuss maximum likelihood estimation. In Section 4, the CHF dataset is considered as an illustration for our proposed model. A simulation study to evaluate the model is presented in Section 5. Finally, in Section 6, main results are discussed.

2. MODEL FORMULATION

The destructive model introduced in Rodrigues *et al.* [11] considers M as a (unobservable) random variable denoting the initial number of concurrent causes that can produce the event of interest with probability mass function (pmf) given by

$$(1)\quad P(M = m; \theta, \phi) = \frac{w(m; \phi)p^*(m; \theta)}{E_\theta[w(M; \phi)]}, \quad m = 0, 1, 2, \dots,$$

where $w(\cdot; \phi)$ is a non-negative weight function indexed by the parameter ϕ, $p^*(\cdot; \theta)$ is the pmf of the Poisson distribution with mean $\theta > 0$. The notation $E_\theta[\cdot]$ indicates that the expectation is taken with respect to the variable M following a Poisson distribution with mean θ. Given $M = m$,

Table 1. Special cases of the PS($\theta, A(\theta)$) distribution, $\theta \in \Theta$. For Binomial and Negative Binomial distributions q is considered known

Distribution	a_m	$A(\theta)$	$A'(\theta)$	Θ
Poisson(θ)	$(m!)^{-1}$	e^{θ}	e^{θ}	$(0, \infty)$
Logarithmic(θ)	$(m+1)^{-1}$	$-\frac{\log(1-\theta)}{\theta}$	$\frac{(1-\theta)\log(1-\theta)+\theta}{\theta^2(1-\theta)}$	$(0, 1)$
Negative Binomial(q, θ)	$\binom{m+q-1}{m}$	$(1-\theta)^{-q}$	$q(1-\theta)^{-(q+1)}$	$(0, 1)$
Binomial(q, θ)	$\binom{q}{m}$	$(1+\theta)^{q}$	$q(1+\theta)^{q-1}$	$(0, \infty)$

let ϱ_j, $j = 1, 2, \ldots, m$, be independent and identically distributed Bernoulli random variables. If the j-th cause produces the event, $\varrho_j = 1$; otherwise, $\varrho_j = 0$. Therefore, for $P(\varrho_j = 1) = p$, the traditional models of Berkson and Gage [1] and Chen *et al.* [2] can be seen as particular cases of $p = 1$ (which implies $M = D$).

The unobserved quantity

$$D = \begin{cases} \varrho_1 + \cdots + \varrho_M, & \text{if } M > 0, \\ 0, & \text{if } M = 0, \end{cases}$$

with $D \leq M$, is the total concurrent causes not destroyed. Clearly, $D \mid M = m \sim \text{Bin}(m, p)$ if $m > 0$ and $P(D = 0 \mid M = 0) = 1$. Also,

$$P(D = d; \theta, p, \phi) = \frac{e^{-\theta p}(\theta p)^d}{d! E_{\theta}[w(M; \phi)]} E_{\theta(1-p)}[w(M + d; \phi)].$$

Differently from the authors, we propose to assume for M a different class of discrete models called the power series distribution (Noack [17]) with pmf given by

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

where $a_m \geq 0$, $\theta > 0$ and $A(\theta) = \sum_{m=0}^{\infty} a_m \theta^m$ is the series function. We denote the distribution in (2) as PS($\theta, A(\theta)$). The main reasons for the choice of this class of models are: *(i)* up to this moment, in this context the PS distribution has not been considered; *(ii)* many popular distributions belong to this class, such as Poisson, logarithmic, negative binomial, among others and; *(iii)* the probability generating function (pgf) of the model have a closer form, which is very relevant for the computation of the population survival function as we will see in a future section.

The first moment for this distribution is $\mathbb{E}(M) = \theta \frac{\partial \log A(\theta)}{\partial \theta}$ and the k-th moment can be computed using the recursive formula

$$\mathbb{E}(M^{k+1}) = \theta \frac{\partial \mathbb{E}(M^k)}{\partial \theta} + \mathbb{E}(M)\mathbb{E}(M^k), \qquad k = 1, 2, \ldots.$$

Depending on a_m, some popular distributions are obtained, as shown in Table 1.

Considering $D \mid M = m \sim Bin(m, p)$ if $m > 0$ with $P(D = 0 \mid M = 0) = 1$, the corresponding marginal distribution is given by the following proposition.

Proposition 1. *For the initial number of causes M, statistically described by the distribution in (2), the total number D of actual concurrent causes will have pmf*

$$(3) \qquad P(D = d; \theta, p) = \frac{(\theta p)^d}{d! A(\theta)} \frac{\partial^d [A(u)]}{\partial u^d}\Big|_{u=\theta(1-p)}.$$

Proof. By the law of total probability,

$$\begin{aligned} P(D = d; \theta, p) &= \sum_{m=d}^{\infty} P(D = d \mid M = m; p) P(M = m; \theta) \\ &= \sum_{m=d}^{\infty} \binom{m}{d} p^d (1-p)^{m-d} \frac{a_m \theta^m}{A(\theta)} \\ &= \frac{(\theta p)^d}{d! A(\theta)} \sum_{m=d}^{\infty} \frac{m!}{(m-d)!} a_m [\theta(1-p)]^{m-d}. \end{aligned}$$

Note that $A(u) = \sum_{m=0}^{\infty} a_m u^m$, so differentiating d times with respect to u, we have that $\frac{\partial^d A(u)}{\partial u^d} = \sum_{m=d}^{\infty} \frac{m!}{(m-d)!} a_m u^{m-d}$. The result follows considering $u = \theta(1-p)$. □

It is straightforward to prove that $\mathbb{E}(D) = \theta p \frac{\partial \log A(\theta)}{\partial \theta}$. In long-term survival models, the pgf has a very important role, because the population survival function can be expressed in terms of that function Rodrigues *et al.* [3]. It can be verified that the pgf for D is given by

$$(4) \qquad \psi_D(s) = \frac{A(\theta(1 - p(1-s)))}{A(\theta)}, \quad s \in (0, 1).$$

Similar to Rodrigues *et al.* [11], we further assume that W_a, $a = 1, \ldots, D$ have common survival function given by $S(\cdot; \lambda) = 1 - F(\cdot; \lambda)$. In addition, if $W_1, \ldots, W_D$ are independent of D, it follows that the (population) survival function based on the FA scheme is $S_{pop}^{FA}(t) = \psi_D(S(t; \lambda))$ whereas, for the LA scheme, $S_{pop}^{LA}(t) = 1 + \psi_D(0) - \psi_D(F(t; \lambda))$. The

Table 2. Population survival and density functions for the three considered schemes

Scheme	$S_{pop}(t)$	$f_{pop}(t)$
FA	$\frac{A(\theta(1-pF(t;\lambda)))}{A(\theta)}$	$\frac{A'(\theta(1-pF(t;\lambda)))}{A(\theta)}\theta p f(t;\lambda)$
LA	$1+\frac{A(\theta(1-p))}{A(\theta)}-\frac{A(\theta(1-pS(t;\lambda)))}{A(\theta)}$	$\frac{A'(\theta(1-pS(t;\lambda)))}{A(\theta)}\theta p f(t;\lambda)$
Mix	$\pi S_{pop}^{FA}(t)+(1-\pi)S_{pop}^{LA}(t)$	$\pi f_{pop}^{FA}(t)+(1-\pi)f_{pop}^{LA}(t)$

Table 3. Survival and density functions related to susceptible individuals for each activation scheme

Scheme	$S_{sus}(t)$	$f_{sus}(t)$
FA	$\frac{A(\theta(1-pF(t;\lambda)))-A(\theta(1-p))}{A(\theta)-A(\theta(1-p))}$	$\frac{A'(\theta(1-pF(t;\lambda)))}{A(\theta)-A(\theta(1-p))}\theta p f(t;\lambda)$
LA	$\frac{A(\theta)-A(\theta(1-pS(t;\lambda)))}{A(\theta)-A(\theta(1-p))}$	$\frac{A'(\theta(1-pS(t;\lambda)))}{A(\theta)-A(\theta(1-p))}\theta p f(t;\lambda)$
Mix	$\pi S_{sus}^{FA}(t)+(1-\pi)S_{sus}^{LA}(t)$	$\pi f_{sus}^{FA}(t)+(1-\pi)f_{sus}^{LA}(t)$

mixture between FA and LA schemes will have survival function given by

$$S_{pop}^{Mix}(t)=\pi S_{pop}^{FA}(t)+(1-\pi)S_{pop}^{LA}(t),$$

with $0\leq\pi\leq 1$. Table 2 shows population survival and density functions (S_{pop} and f_{pop}, respectively) for the proposed model.

Henceforth, we call our model destructive power series (DPS) cure rate model and, for each particular activation schemes in Table 2, we denote DPS-FA, DPS-LA and DPS-Mix, respectively. When using a specific distribution belonging to the power series model, we denote the models as DP (Destructive Poisson), DL (Destructive logarithmic), DNB (Destructive Negative binomial) and DBin (Destructive binomial), appending the corresponding activation scheme. For instance, DP-FA, DL-LA, DNB-Mix, etc.

Note that the cure probabilities for the models in Table 2 are all equal to

$$q_0=A(\theta(1-p))/A(\theta).$$

As for the distribution of latent times W_a, one can assume any parametric model. In this paper, we will consider the Weibull distribution, given its widespread use and suitability in many biological problems; however, other models could be considered. The survival function for W_a is then given by $S(t;\lambda)=\exp\{-e^{\alpha}t^{\nu}\}$, where $\lambda=(\alpha,\nu)$, $\alpha\in\mathbb{R}$ and $\nu,t>0$. The survival and density functions for susceptible individuals (when $D\geq 1$) will be denoted by $S_{sus}(t)$ and $f_{sus}(t)$, respectively, and are presented in Table 3. It is straightforward to show that they are proper survival functions, as expected.

3. ESTIMATION

Subject to right censoring, the actual observable data will be associated to random variables $T_i=\min(T_i^*,C_i)$ and $\delta_i=I(T_i^*\leq C_i)$, $i=1,\ldots,n$, where T_i^* and C_i denote failure and censoring times, respectively, and δ_i corresponds to the failure indicator. In addition, we associate to the i-th individual, $i=1,\ldots,n$, a set of covariates $z_{1i}=(1,z_{1i1},\ldots,z_{1ir_1})$ related to the initial number of causes and $z_{2i}=(1,z_{2i1},\ldots,z_{2ir_2})$ related to the activation probabilities for non-destroyed cells in such a way that

$$\log\theta_i=z_{1i}^{\top}\beta_1 \qquad \text{and} \qquad \log\left(\frac{p_i}{1-p_i}\right)=z_{2i}^{\top}\beta_2,$$

where β_1 and β_2 are vectors of unknown parameters with dimensions (r_1+1) and (r_2+1), respectively. In order to avoid identifiability issues in the sense of Li *et al.* [18], we have to consider z_1 and z_2 as not sharing common elements. Specifically for the NB model, it is needed not to have the intercept in one of the term parameter vectors. From a practical point of view, it seems more appropriate to take β_2 without intercept, as this will imply that, for an individual with all covariates equal to zero, the probability of activation for the initial cells is 0.5, which seems natural if the covariates are centered (for the continuous case, for instance).

Maximum likelihood estimators for ψ, say $\widehat{\psi}$, are obtained by maximizing the log-likelihood function for $\psi=(\beta_1,\beta_2,\lambda,\pi)$, given by

$$(5) \quad \ell(\psi)=\sum_{i=1}^{n}\left[\delta_i\log f_{pop}(t_i)+(1-\delta_i)\log S_{pop}(t_i)\right],$$

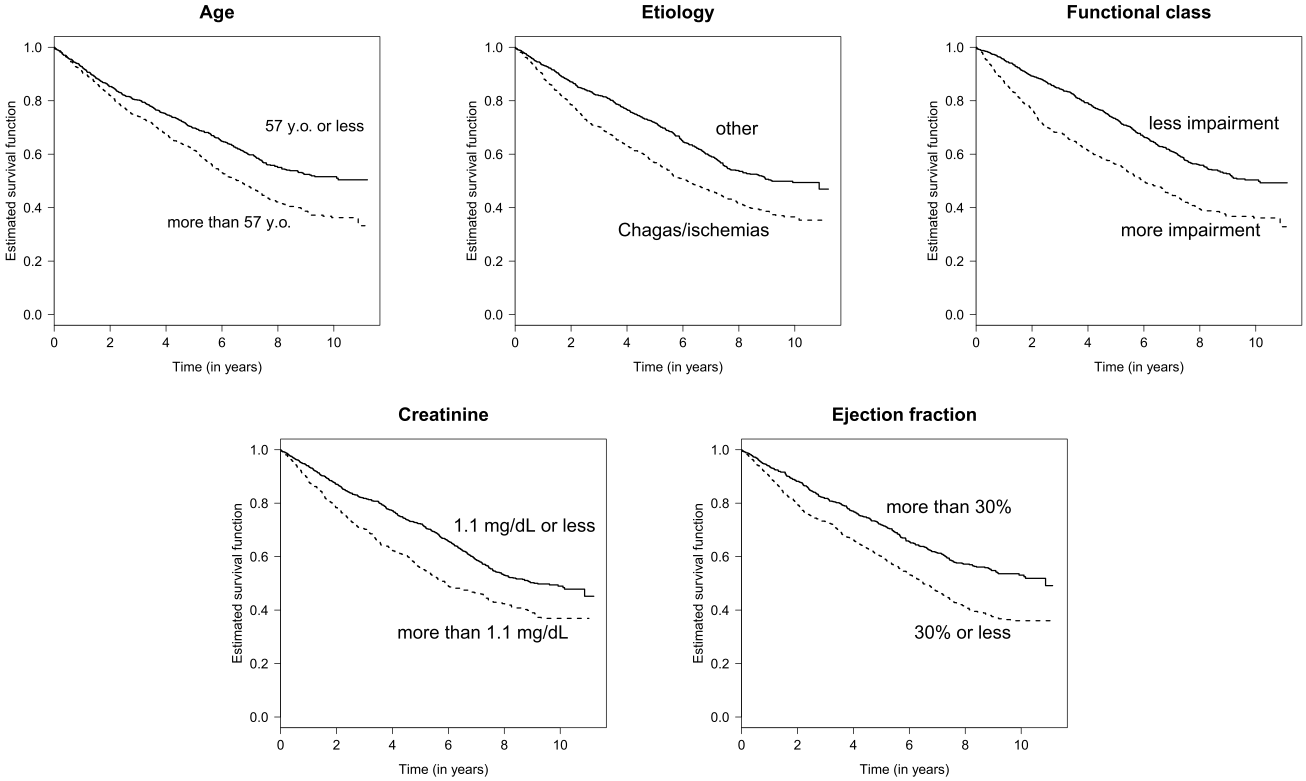

Figure 2. Kaplan-Meier estimates for each covariate in the CHF dataset.

where $f_{pop}(\cdot)$ and $S_{pop}(\cdot)$ are given in Table 2 and δ_i denotes the failure indicator for the i-th individual, $i = 1, \ldots, n$. Standard errors are estimated based on the Hessian matrix.

In practice, the selection of covariates for θ and p may be a problem. Up to now very few procedures have been proposed to deal with this issue. [11] consider all possible combinations of covariates and select the model which provides a smaller AIC. However, in their problem the data set has only 2 covariates, producing a reasonable number of combinations. Our CHF dataset, which is actually a subset of the original data, we have 5 covariates, considerably increasing the number of combinations, even when the intercept is included in θ. At the end, we considered the AIC criteria based on 180 model combinations.

4. APPLICATION TO HEART FAILURE DATA

In this section, we apply the destructive power series model to the CHF dataset. A total of 2,128 patients treated in the Heart Institute from the Medical School of the University of São Paulo, Brazil were followed from July 2003 to April 2015. The time to failure is defined as the number of years from enrolment in the hospital's protocol until death, loss of follow-up or end of study. The mean and median follow-up times were 5.23 and 4.84 years, respectively (sd=3.44). The following subset of the original covariates were selected for this illustration:

- *age*: age of the patient, in years, at the time of enrolment in the study (mean = 56.86, median=57, sd = 3.44);
- *eti*: dummy variable indicating if the cause of CHF was associated to ischemic cardiomyopathy or Chagas disease (1); or other causes (0) ($n = 644$ and $n = 969$, respectively);
- *fcl*: dummy variable indicating if patient was in functional classes of more (1), or less (0) impairment related to the CHF ($n = 695$ and $n = 918$, respectively);
- *ejf*: left ventricle ejection fraction (in %, mean=33.47, median=30, sd=11.30);
- *cre*: level of serum creatinine at the enrolment time (in mg/dL, mean=1.21, median=1.1, sd=0.85).

Covariates *age*, *cre* and *ejf* were centered at 57 years, 1.1 mg/dL and 30%, respectively. Those values correspond to the respective median values. The data is characterized by a considerable amount of missing data. For this illustration, we considered only subjects with complete data for all described covariates, resulting in 1,613 patients, with 900 of them presenting to censored times (56% of total). An initial analysis suggested the combination of etiologies described above. Despite the long follow-up time, the Kaplan-Meier estimators (KM) for each covariate in Figure 2 show

Table 4. Models with lowest AIC among all combinations under identifiability for the CHF dataset

Distribution of M	Scheme activation	Covariates in z_1	Covariates in z_2	Log-likelihood	Number of parameters	AIC
Poisson	Mix	*age, eti, fcl, ejf*	*cre*	-2378.152	9	4774.304
Poisson	Mix	*age, fcl, ejf*	*eti, cre*	-2379.479	9	4776.958
Poisson	LA	*age, eti, fcl, ejf*	*cre*	-2380.231	8	4776.462
NB ($q = 3$)	Mix	*age, eti, fcl, ejf*	*cre*	-2379.867	9	4776.734
NB ($q = 2$)	LA	*age, eti, fcl, ejf*	*cre*	-2380.416	8	4776.832

Table 5. Estimated parameters for DP-Mix model in the heart failure dataset

			Confidence interval 95%			
Parameter	Estimate	s.e.	Lower limit	Upper limit	exp(Estimate)	p-value
β_0	1.2211	0.3384	0.5578	1.8844	–	0.0003
β_{1age}	0.0182	0.0036	0.0111	0.0253	1.0184	<0.0001
β_{1eti}	0.4954	0.0933	0.3125	0.6783	1.6412	<0.0001
β_{1fcl}	0.5404	0.0942	0.3557	0.7251	1.7166	<0.0001
β_{1ejf}	-0.0228	0.0043	-0.0312	-0.0144	0.9774	<0.0001
β_{2cre}	1.8664	0.3587	1.1634	2.5693	–	<0.0001
π	0.8791	0.0416	0.7976	0.9607	–	–
α	-3.4727	0.3293	-4.1182	-2.8273	–	–
ν	1.1949	0.0507	1.0955	1.2943	–	–

a somewhat large proportion of censored patients, with a plateau towards the end of the study. This feature suggests the presence of long-term survivors, with lifetimes greater than expected, eventually similar to the general (non-CHF) population.

The DPS model considering the three activation schemes discussed in Section 2 were fitted. All combinations of the covariates age, etiology, functional class, creatinine and ejection fraction were included. If any problem related to identifiability was detected, the corresponding combination was discarded. Models with smaller AIC are presented in Table 4, where we can conclude that age, etiology, functional class and ejection fraction are related to the number of initial concurrent causes whereas creatinine is related to the activation probability p. The Poisson distribution in combination with the mixture activation scheme seems to have the best performance. Estimates related to this model are presented in Table 5. Note that, considering a 5% level of significance for the Wald test, all covariates are statistically significant. Figure 3 illustrates the influence of each covariate on the estimated survival function. Finally, since the parameter $\theta_i = \exp\left(z_{1i}^{\top}\beta\right)$ in the Poisson model represents the mean of number of initial causes, we may conclude the following:

- For each additional year of age at enrolment, the number of initial causes of death increases, in average, by 1.9% [95% confidence interval = (1.1%-2.6%)].
- The number of initial causes of death for CHF patients with ischemic cardiomyopathy or Chagas disease is, in average, 64.1% larger than for patients with CHF caused by other conditions [95% CI=(36.7%-97.1%)].
- The expected number of initial causes related to the heart failure in patients with more impairment is increased by 71.7% when compared to less impairment patients [95% CI=(42.7%-106.5%)].
- Increasing the ejection fraction in 1% decreases the expected number of causes of death related to the heart failure in 2.3% [95% CI = (1.5%-3.2%)].
- The activation probabilities p_i are estimated as 0.72, 0.89 and 0.98 for patients with creatinine levels equal to 0.5, 1.1 and 2.0 mg/dL, respectively.
- Estimate from the mixing coefficient π shows that 87.91% [95% CI=(79.76%-96.07%)] have a survival experience in the lifespan for CHF patients whereas 12.09% [95% CI = (3.93%-20.24%)] of them would have a longer than expected survival time (though inferior to the general non-CHF population).
- Figure 4 shows the behaviour of estimates for the proportion of long-term patients, considering different ages and values of ejection fraction. As expected, it is less likely to have long-term patients (similar to the general population) as age increases. The reverse effect is observed for ejection fraction.

5. SIMULATION STUDIES

In this section we present two simulation studies. The first is related to verify the performance of the mle in DP-FA, DP-LA, DP-mix, DNB-FA, DNB-LA and DNB-mix models in finite sample when the model is well specified. The second is devoted to study the performance of the mle estimators in DP-mix and DNB-mix when the model is misspecified.

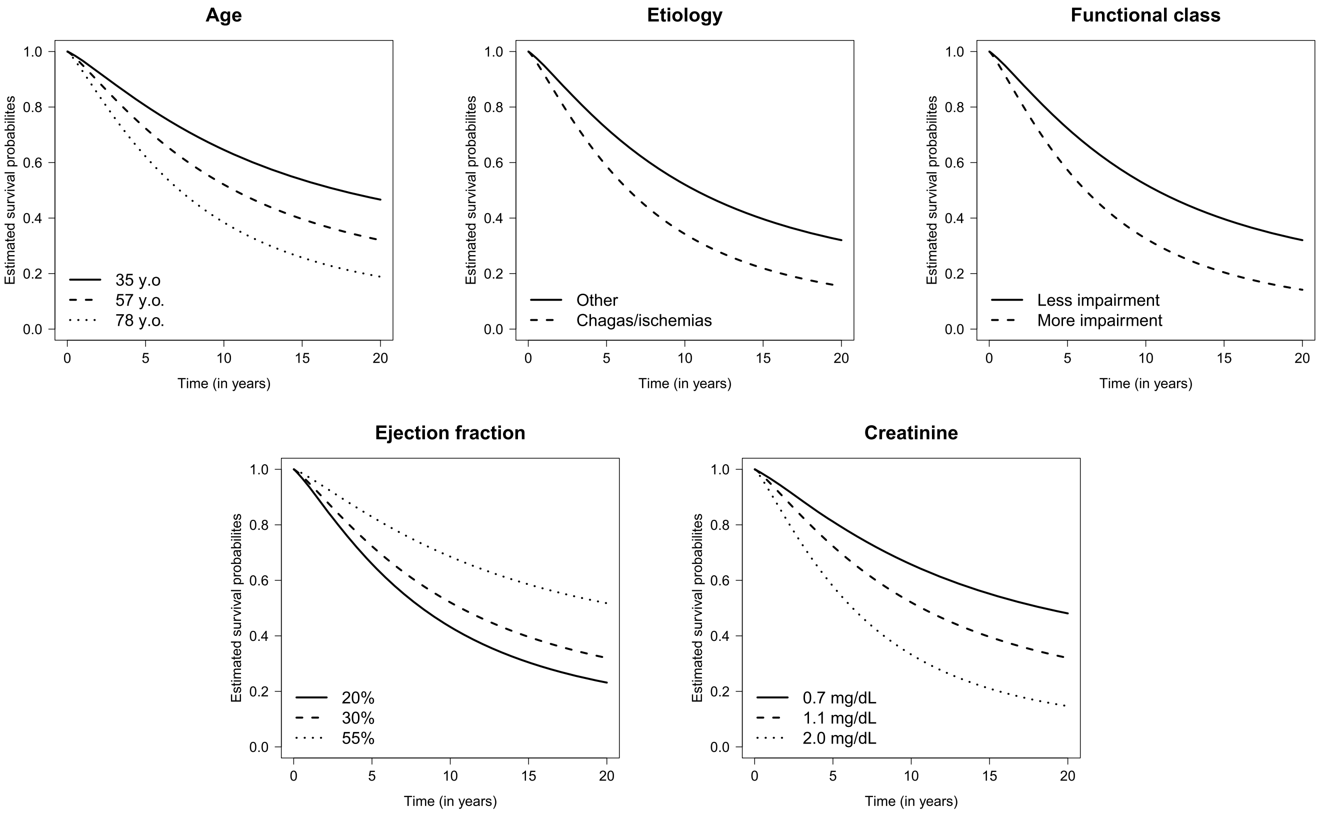

Figure 3. Estimated survival function for DP-mix model in the CHF dataset. When not varying, covariates were taken as: age= 57 years old, etiology= Other, functional class= Less impairment, ejection fraction= 30% and creatinine= 1.1 mg/dL.

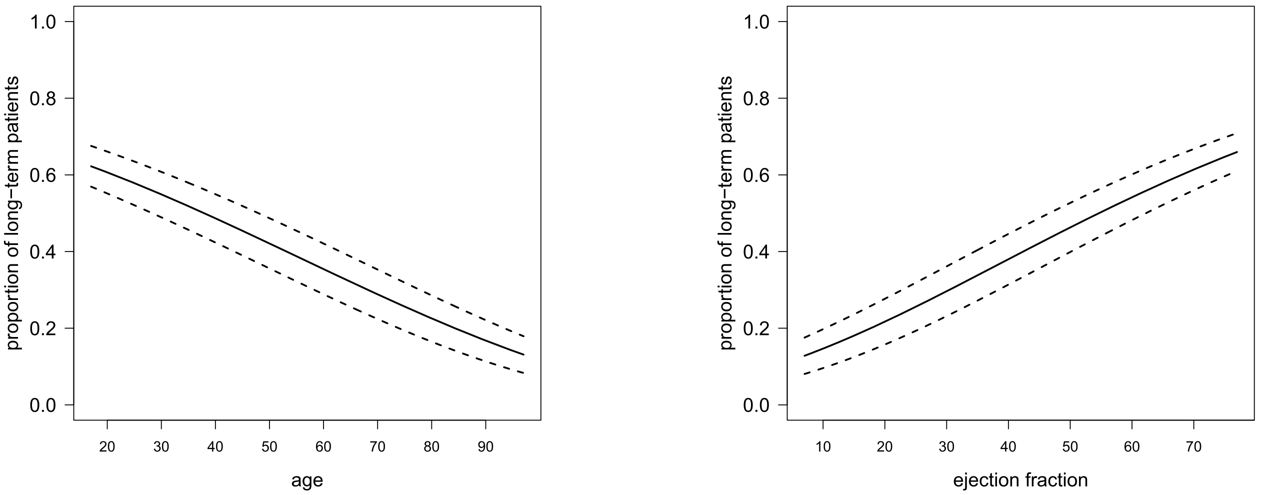

Figure 4. Proportion of long-term patients and corresponding 95% pointwise confidence interval for: (a) patients with etiology other than Chagas/ischaemia, less impairment, creatinine equals to 0.8 mg/dL and ejection fraction of 40% and different ages (left panel) and (b) 60 year old patients, with Chagas disease, more impairment, creatinine equals to 2 mg/dL and different levels of ejection fraction (right panel).

Table 6. Simulation study for DP-mix and DNB-mix models with different sample sizes

		DP-Mix			DNB-Mix		
n	Parameter	Bias	MSE	CP (95%)	Bias	MSE	CP (95%)
100	β_{1age}	0.003	0.000	0.918	0.005	0.001	0.936
	β_{1eti}	0.056	0.237	0.917	0.071	0.309	0.940
	β_{1fcl}	0.055	0.230	0.918	0.135	0.317	0.940
	β_{1ejf}	-0.002	0.000	0.909	-0.007	0.001	0.938
	β_{2cre}	0.205	0.576	0.942	0.254	0.853	0.944
	α	0.235	0.310	0.912	0.264	0.349	0.923
	ν	0.198	0.412	0.920	0.207	0.361	0.919
	π	-0.004	0.020	0.526	-0.019	0.011	0.321
250	β_{1age}	0.001	0.000	0.936	0.004	0.000	0.939
	β_{1eti}	0.016	0.067	0.938	0.063	0.090	0.949
	β_{1fcl}	0.016	0.067	0.941	0.085	0.094	0.944
	β_{1ejf}	-0.001	0.000	0.934	-0.004	0.000	0.944
	β_{2cre}	0.072	0.151	0.946	0.136	0.252	0.947
	α	0.031	0.085	0.935	0.035	0.089	0.938
	ν	0.021	0.101	0.929	0.027	0.075	0.938
	π	0.009	0.010	0.740	0.019	0.002	0.619
500	β_{1age}	0.000	0.000	0.947	0.003	0.000	0.939
	β_{1eti}	0.005	0.031	0.944	0.054	0.043	0.949
	β_{1fcl}	0.009	0.030	0.947	0.067	0.047	0.944
	β_{1ejf}	-0.000	0.000	0.945	-0.004	0.000	0.928
	β_{2cre}	0.026	0.066	0.950	0.084	0.114	0.946
	α	0.010	0.042	0.942	0.012	0.035	0.940
	ν	0.008	0.039	0.939	0.015	0.042	0.941
	π	0.010	0.005	0.863	0.018	0.001	0.847
1000	β_{1age}	0.000	0.000	0.948	0.003	0.000	0.944
	β_{1eti}	0.004	0.015	0.949	0.018	0.022	0.950
	β_{1fcl}	0.006	0.015	0.950	0.048	0.026	0.949
	β_{1ejf}	-0.000	0.000	0.948	-0.004	0.000	0.917
	β_{2cre}	0.014	0.031	0.950	0.039	0.066	0.948
	α	0.009	0.021	0.947	0.014	0.030	0.948
	ν	0.005	0.015	0.948	0.015	0.025	0.947
	π	0.006	0.003	0.929	0.012	0.001	0.914

5.1 Recovery parameters

In this section we present a simulation study to assess the performance of the estimation procedure of Section 3. We follow a similar structure for the covariates motivated by the CHF dataset. Continuous covariates were also centered. For the concurrent causes, we consider the Poisson and negative binomial models. We also consider the three activation schemes (FA, LA and mix) discussed in the previous sections.

Given the highly skewed nature of age, this covariate was drawn from the skew-normal distribution. Creatinine was drawn from normal distribution and ejection fraction was drawn from a uniform distribution. The values for the binary covariates etiology and functional class were drawn from Bernoulli distributions with success probabilities 0.60 and 0.57, respectively. The mean, variance and shape parameters for the skew-normal model was chosen based on the sample values observed for the variables age and creatinine in the CHF dataset, i.e., skew-normal$(-0.125, 12.895, 0.951)$ (see Fernández and Steel [19] for details about the parametrization used for the skew-normal model). Similarly, the mean and variance for creatinine and parameters related to the uniform distribution for ejection fraction were computed based on sample values observed in CHF the dataset. Same procedure, based on the Bernoulli distribution, was adopted in the allocation of subjects for each etiology and functional class.

For each patient, we considered the same combination of covariates associated to z_1 and z_2 (see Table 4). M_i was drawn from the Poisson or NB model depending on the particular case. For $M_i = 0$, we defined $D_i = 0$. Given $M_i > 0$, we simulated D_i from the conditional Binomial distribution with parameters M_i and p_i. If $D_i = 0$, we assigned the respective failure time as $+\infty$; otherwise, we simulated lifetimes $W_1, \dots, W_{D_i}$ from a Weibull distribution. For the FA and LA schemes, failure times are considered as the minimum or the maximum among those times, respectively; for the mix scheme, the failure time is considered as a convex combination of the minimum or maximum

Table 7. Sensitivity analysis for DP-mix and DNB-mix models (FA is the true activation scheme)

		DP-FA			DNB-FA		
n	Parameter	Bias	MSE	CP (95%)	Bias	MSE	CP (95%)
100	β_{1age}	0.005	0.007	0.924	0.013	0.003	0.936
	β_{1eti}	0.061	0.254	0.919	0.067	0.309	0.939
	β_{1fcl}	0.049	0.221	0.908	0.128	0.315	0.934
	β_{1ejf}	-0.011	0.003	0.914	-0.010	0.000	0.943
	β_{2cre}	0.146	0.599	0.970	0.156	0.830	0.931
	α	0.181	0.333	0.866	0.271	0.410	0.929
	ν	0.228	0.375	0.957	0.224	0.336	0.928
	π	-0.072	0.028	0.529	-0.023	0.012	0.329
250	β_{1age}	-0.004	0.003	0.941	0.008	0.002	0.939
	β_{1eti}	0.016	0.068	0.927	0.066	0.088	0.942
	β_{1fcl}	0.028	0.069	0.927	0.087	0.088	0.937
	β_{1ejf}	-0.001	0.003	0.936	-0.008	0.005	0.946
	β_{2cre}	0.071	0.173	0.963	0.131	0.233	0.939
	α	0.034	0.062	0.931	0.051	0.135	0.935
	ν	0.033	0.120	0.955	0.029	0.102	0.939
	π	-0.050	0.010	0.741	-0.022	0.006	0.621
500	β_{1age}	-0.004	0.001	0.952	0.005	0.002	0.941
	β_{1eti}	0.011	0.028	0.952	0.046	0.044	0.948
	β_{1fcl}	0.004	0.031	0.941	0.070	0.040	0.945
	β_{1ejf}	0.002	0.001	0.944	-0.004	0.001	0.947
	β_{2cre}	0.027	0.111	0.959	0.091	0.154	0.941
	α	0.016	0.047	0.936	0.024	0.024	0.941
	ν	0.007	0.027	0.951	-0.005	0.024	0.941
	π	-0.038	0.008	0.867	-0.021	0.002	0.852
1000	β_{1age}	0.002	0.001	0.950	0.002	0.001	0.945
	β_{1eti}	0.002	0.011	0.949	0.022	0.023	0.949
	β_{1fcl}	0.003	0.012	0.946	0.047	0.025	0.947
	β_{1ejf}	0.000	0.000	0.946	-0.002	0.002	0.949
	β_{2cre}	0.026	0.020	0.954	0.026	0.054	0.958
	α	0.004	0.036	0.960	0.009	0.013	0.946
	ν	0.002	0.013	0.950	0.025	0.030	0.947
	π	-0.016	0.005	0.931	-0.013	0.002	0.926

values, with coefficients π and $1-\pi$, respectively. Additionally, the lifetimes were censored at a value c simulated from uniform$(0, 11.2)$, i.e., based on the maximum time observed in the CHF dataset. The percentage of censoring ranged from 49% to 75% (average of 62%).

In order to assess the behaviour of the estimators in finite samples we also considered four sample sizes $(100, 250, 500, 1000)$. Each case was replicated 1,000 times. Bias, mean squared error (MSE) and the 95%-coverage probability (CP) for the mix activation scheme are shown in Table 6.

Note that for both models, the bias and the MSE related to the covariates decrease as the sample size increases. Moreover, the coverage probabilities (CP) are closer to the nominal value for larger values of n. These results suggest that the estimator for β coefficients and the estimators related to the time-to-event α and ν are asymptotically consistent; indeed, they are well estimated even when the sample sizes are moderate (e.g., $n = 250$). The CP's related to π performed fairly well for the case $n = 1{,}000$, suggesting that to construct a reasonable confidence interval for π it may be necessary a large sample size, as is the case for the CHF dataset. Similar results were obtained for the FA and LA activation schemes, but we omit such tables here.

5.2 Misspecification of the activation scheme

This simulation study is devoted to study the performance of the mle in the DNB-mix model if the activation scheme is FA or LA. We consider the same structure to draw the data and the same values for parameters considered in last study, except for π, where we consider $\pi = 1$ and $\pi = 0$ for FA and LA schemes, respectively. We also consider sample size of 100, 250, 500 and 1,000. Results summarized using the bias, MSE and coverage probabilities are presented in Tables 7 and 8. Note that conclusions for the mle of the components of the vector β, α and ν are similar than last study, i.e., such behaviour is reasonable in terms of bias, MSE and CP. On the other hand, the bias of the estimator of π can be considerable for not so large sample sizes, but more serious is, again, confidence intervals covering less times that

Table 8. *Sensitivity analysis for DP-mix and DNB-mix models (LA is the true activation scheme)*

		DP-LA			DNB-LA		
n	Parameter	Bias	MSE	CP (95%)	Bias	MSE	CP (95%)
100	β_{1age}	0.012	0.009	0.916	-0.004	0.005	0.938
	β_{1eti}	0.070	0.247	0.932	0.086	0.308	0.936
	β_{1fcl}	0.073	0.218	0.918	0.133	0.323	0.935
	β_{1ejf}	0.015	0.015	0.912	-0.009	0.005	0.926
	β_{2cre}	0.284	0.580	0.973	0.173	0.929	0.978
	α	0.289	0.254	0.913	0.316	0.408	0.918
	ν	0.185	0.433	0.919	0.249	0.359	0.935
	π	0.021	0.028	0.531	0.029	0.017	0.324
250	β_{1age}	0.008	0.008	0.929	-0.001	0.005	0.939
	β_{1eti}	0.011	0.056	0.938	0.062	0.082	0.939
	β_{1fcl}	0.024	0.063	0.938	0.098	0.109	0.939
	β_{1ejf}	0.009	0.007	0.930	-0.008	0.004	0.938
	β_{2cre}	0.046	0.210	0.961	0.130	0.233	0.881
	α	0.019	0.077	0.934	0.066	0.065	0.929
	ν	0.042	0.072	0.921	0.060	0.087	0.936
	π	0.015	0.013	0.747	0.023	0.006	0.619
500	β_{1age}	0.005	0.004	0.948	0.001	0.004	0.942
	β_{1eti}	0.009	0.022	0.943	0.059	0.039	0.941
	β_{1fcl}	0.017	0.036	0.944	0.069	0.048	0.942
	β_{1ejf}	0.005	0.003	0.949	-0.005	0.002	0.940
	β_{2cre}	0.038	0.083	0.957	0.055	0.139	0.983
	α	0.012	0.010	0.943	0.025	0.067	0.935
	ν	0.012	0.037	0.935	0.034	0.060	0.941
	π	0.012	0.009	0.863	0.020	0.006	0.849
1000	β_{1age}	0.001	0.002	0.948	0.002	0.002	0.945
	β_{1eti}	0.002	0.017	0.945	0.020	0.021	0.949
	β_{1fcl}	0.005	0.013	0.948	0.050	0.025	0.947
	β_{1ejf}	0.002	0.001	0.948	-0.002	0.001	0.945
	β_{2cre}	0.006	0.055	0.952	0.040	0.067	0.940
	α	0.009	0.006	0.949	0.012	0.023	0.945
	ν	0.004	0.004	0.945	0.025	0.018	0.946
	π	0.008	0.004	0.931	0.014	0.001	0.915

expected the true value of π. Our recommendation is consider, to obtain a reliable estimator to this parameter, a large sample size (say $n \geq 1000$).

6. FINAL DISCUSSION

We proposed a new cure rate model, motivated by a real data set related to patients with congestive heart failure. The model considers that a fraction of patients may have the same lifetime pattern as the general population, i.e., subjects without the disease or *cured*. Also incorporates the possibility that, among susceptible individuals, there may exist a proportion (say π) of them living *more than expected*. The particular cases $\pi = 1$ and $\pi = 0$ lead to schemes known as *first* and *last* activation schemes.

Estimation is easily performed based on the maximum likelihood method, as long as some care is taken with respect to identifiability. Simulation studies shows that estimation of parameters in the new model has, in general, good performance for finite samples, except for the parameter π, when a larger sample size ($n > 1000$ for our considered cases) may be needed to achieve reasonable coverage probability, although bias and mean square error performance behave fairly well. However, since the cure probability does not depend on π, it will always be adequately estimated, regardless the sample size. When choosing the probabilistic model for the number of latent factors, it is important to take into consideration the available sample size, as shown in the simulation: the Negative binomial model requires a larger number of subjects when compared to the Poisson model. Results observed in a read data set application were consistent with what is observed in daily medical practice.

ACKNOWLEDGEMENTS

The first author has been partially supported by FONDECYT (Fondo Nacional de Desarrollo Científico y Tecnológico) Grant 11160670.

Received 2 January 2019

REFERENCES

[1] Berkson J, Gage R. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952: **47**: 501–515.

[2] Chen MH, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 1999: **94**: 909–919. MR1723307

[3] Rodrigues J, Cancho VG, de Castro MA, Louzada-Neto F. On the unification of the long-term survival models. *Statistics and Probability Letters* 2009: **79**: 753–759. MR2662300

[4] Rodrigues J, de Castro MA, Cancho VG, Balakrishnan N. COM-Poisson cure rate survival model and an application to a cutaneous melanoma data. *Journal of Planning and Inference* 2009: **139**: 3605–3611. MR2549108

[5] Cancho VG, Louzada F, Barriga GDC. The Geometric Birnbaum-Saunders regression model with cure rate. *Journal of Statistical Planning and Inference* 2013: **142**: 993–1000. MR2863887

[6] Cancho VG, Louzada F, Ortega EM. The power series cure rate model: an application to a cutaneous melanoma data. *Communications in Statistics – Simulation and Computation* 2013: **42**: 586–602. MR3020088

[7] Ortega EMM, Barriga GDC, Hashimoto EM, Cancho VG, Cordeiro GM. A new class of survival regression models with cure fraction. *Journal of Data Science* 2014: **12**: 107–136. MR3099500

[8] Ortega EMM, Cordeiro GM, Campelo AK, Kattan MW, Cancho VG. A power series beta Weibull regression model for predicting breast carcinoma. *Statistics in Medicine* 2015: **34**: 1366–1388. MR3322774

[9] Cordeiro GM, Cancho VG, Ortega EM, Barriga GDC. A model with long-term survivors: Negative binomial Birnbaum-Saunders. *Communication in Statistics – Simulation and Computation* 2016: **5**: 1370–1387. MR3462152

[10] Gallardo DI, Gómez HW, Bolfarine H. A new cure rate model based on the Yule-Simon distribution with application to a melanoma data set. *Journal of Applied Statistics* 2016: **44**: 1153–1164. MR3638437

[11] Rodrigues J, de Castro M, Balakrishnan N, Cancho VG. Destructive weighted Poisson cure rate models. *Lifetime Data Analysis* 2011: **17**: 333–346. MR2806936

[12] Yang G, Chen C. A stochastic two-stage carcinogenesis model: a new approach to computing the probability of observing tumour in animal bioassays. *Mathematical Biosciences* 1991: **104**: 247–258. MR1102839

[13] Cooner F, Banerjee S, Carlin BP, Sinha D. Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association* 2007: **102**: 560–572. MR2370853

[14] Cancho VG, de Castro M, Dey DK. Long-term survival models with latent activation under a flexible family of distributions. *Brazilian Journal of Probability and Statistics* 2013: **27**: 585–600. MR3105045

[15] Cancho VG, Bandyopadhyay D, Louzada F, Yiqi B. The destructive negative binomial cure rate model with a latent activation scheme. *Statistical Methodology* 2013: **13**: 48–68. MR3036216

[16] Gallardo DI, Bolfarine H, Pedroso-de-Lima, AC. A clustering cure rate model with application to a sealant study. *Journal of Applied Statistics* 2017: **44**: 2949–2962. MR3721083

[17] Noack A. On a class of discrete random variables. *Annals of Mathematical Statistics* 1950: **21**: 127–132. MR0033472

[18] Li CS, Taylor J, Sy J. Identifiability of cure models. *Statistics and Probability Letters* 2001: **54**: 389–395. MR1861384

[19] Fernández C, Steel MFJ. Bayesian regression analysis with scale mixtures of normals. *Econometric Theory* 2000: **16**: 80–101. MR1749020

Diego I. Gallardo
Departamento de Matemática
Facultad de Ingeniería
Universidad de Atacama
Chile
E-mail address: diego.gallardo@uda.cl

Heleno Bolfarine
Institute of Mathematics and Statistics
University of São Paulo
Brazil
E-mail address: hbolfar@ime.usp.br

Antonio C. Pedroso-de-Lima
Institute of Mathematics and Statistics
University of São Paulo
Brazil
E-mail address: acarlos@ime.usp.br

Jose S. Romeo
SHORE and Whariki Research Centre
College of Health
Massey University
New Zealand
E-mail address: j.romeo@massey.ac.nz