Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/150212

The Destructive Cure Rate Regression in Cancer Prognosis and Prediction

Vicente G. Cancho[1], *Gauss M. Cordeiro*[2], *Gladys Barriga* [3], *Edwin M. M. Ortega*[4,*] *and Michael W. Kattan*[5]

[1]ICMC, Universidade de São Paulo, São Carlos, SP, Brazil
[2]Departamento de Estatística, Universidade Federal de Pernambuco, Recife, PE, Brazil
[3]FEB, Universidade Estadual Paulista, Bauru, SP, Brazil
[4]Departamento de Ciências Exatas, ESALQ, Universidade de São Paulo, Piracicaba, SP, Brazil
[5]Department of Quantitative Health Sciences, Cleveland Clinic, Desk JJN3-01, USA

Received: 10 Nov. 2021, Revised: 9 Dec. 2020, Accepted: 7 Feb. 2021
Published online: 1 Mar. 2022

Abstract: A new flexible cure rate survival regression is proposed for predicting cancer prognosis, which provides a more realistic interpretation of the biological mechanism of the event of interest. The new regression predicts breast carcinoma survival in post-mastectomy women but it can be applied to different types of surgery offered to treat cancer.

Keywords: Breast carcinoma, Cure rate regression, Power series family, Survival models

Regression models which study the distribution of lifetimes have wide applicability in cancer research and other chronic disease prognosis analyses, where some explanatory variables in the surgery or pathology report may be associated with outcomes. The event of interest in many survival studies or cancer-relapse trials can be the death of a patient or a tumor recurrence. Nowadays, a high portion of the patients are expected to be *cured* and there exists a vast literature on 'cure rate models' or 'long-term survival models'; for more details see the key references [1] and [2] or the books by [3] and [4]. Further statistical regressions in this area are due to [5], [6], [7], [8], [9], [10], [11], [12], [13] and [14], among others.

Survival data that present a proportion of individuals in the population who are not susceptible to the event of interest are generally modeled considering a reasonable fraction of cured (or survival models with curing rate). Empirically, this feature can be noted in the Kaplan-Meier estimate of the survival function which has a right tail at an approximately constant level and strictly greater than zero for a considerable period. It is often considered that "cure" is related to survival beyond five years.

Incorporating a surviving fraction in survival models, it is almost impossible to verify assumptions related to latent events, even from biological and/or physical point of views. The distributional assumptions are debatable. Accordingly, we consider a general class of distributions, which includes several plausible distributions.

The article defines a general class of destructive survival cure rate models, where the initial number of unobservable event-related competing causes has a power series distribution, and the probability p of a competing cause not being eliminated by the initial treatment is related to a set of covariates. The proposed regression does not have identifiability problems and it can be reduced if $p = 1$ to the survival regression for modeling lifetime data without cure fraction.

The research examines data collected by [15], where there are 284 women treated with mastectomy and axillary lymph node dissection at Memorial Sloan-Kettering Cancer Center (New York) introduced in Section 6. The time-event-time data considered in this study is the time until the patient's death. The left panel of Figure 1 reports the overall Kaplan-Meier survival curve for the treated breast cancer patients. The plateau points out the presence of cure fraction on the patients, where about 75% of patients did not die during the period of study. Some clinical covariates may affect the probability of cure as shown in Figure 1 (right panel), where the cumulative hazard function of death depends on the patient age.

* Corresponding author e-mail: edwin@usp.br

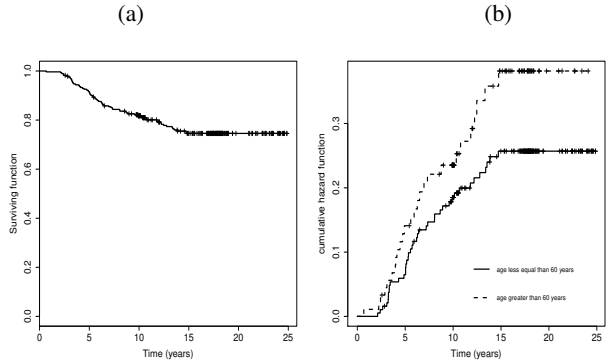

Fig. 1: (a) Kaplan-Meier curves for the breast carcinoma cohort data. (b) Cumulative hazard function stratified by age.

The proposed destructive model can in principle be applied to any pathology, especially in the case of cancer. The initial risk factors are malignant cells and a random variable models the number of live malignant cells that are descendants of a given malignant cell over a period of time. This assumption was not adopted by [9] on predicting the cure and recurrence of breast cancer. In these terms, the new proposal accurately models the total number of live malignant cells at a specific point in time, which is more realistic, since all the variables involved are latent.

The rest of the article is divided into six sections. In Section 1, a new model is formulated for the time distribution of the entire population. In Section 2, some structural properties of the recurrence time for the non-cured individuals are investigated. Likelihood inference is addressed in Section 3. A simulation study in Section 4 reports some statistical properties of the maximum likelihood estimators (MLEs). In Section 5, the methodology is illustrated on a breast cancer data set. Section 6 offers some concluding remarks.

1 The new model

The number of altered cells before an initial treatment is represented by the random variable (rv) M with a power series probability mass function (pmf) [16]

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

where $a_m\geq 0$, $\theta\in(0,s)$ (s can be ∞) is the the power parameter and $A(\theta)=\sum_{m=1}^{\infty}a_m\,\theta^m$ is a finite generator function.

The binomial, Poisson, geometric and logarithmic are four important distributions in the class (1).

Let W_j $(j=1,\ldots,m)$ (for fixed $M=m$) denote the number of living malignant cells that are descendants of the initiated malignant cell j during some period. Consider that $W_1,\cdots,W_m$ are independent and identically distributed (iid) rvs with a Bernoulli distribution (independent of M) with success probability ϕ, which indicates the probability of an undestroyed clonogenic cell. Let $N=W_1+\cdots+W_M$ $(N\leq M)$ be the total number of altered cells among the M initial cells, which are not damaged by the treatment [17]. The probability generating function (pgf) of N is

$$G_N(s)=\frac{A([1-\phi+\phi\,s]\theta)}{A(\theta)},\ 0\leq|s|\leq 1. \tag{2}$$

Let Z_j $(j=1,\ldots,N)$ be the time to the event for the j-th competing cause. Conditional on N, consider that the Z_j's are iid rvs with cumulative distribution function (cdf) $F(t)$ (for $t>0$) which do not depend on N. The total number of competing causes N and the time Z_j are not observable variables. Let $T=\min(Z_1,\cdots,Z_N)$ be the observable time to the event of interest, where $T=\infty$ if $N=0$ and $P(T=\infty|N=0)=1$.

The improper survival function ([18], [7]) (under these conditions) for the entire population is

$$S_{\rm pop}(t)=\frac{A(\theta[1-\phi F(t)])}{A(\theta)}. \tag{3}$$

The cured probability $p_0=\lim_{t\to\infty}S_{\rm pop}(t)$ follows from (3)

$$p_0=\frac{A(\theta[1-\phi])}{A(\theta)}>0. \tag{4}$$

The density function associated with (3) is

$$f_{\rm pop}(t)=-S'_{\rm pop}(t)=\frac{A'(\theta\,[1-\phi F(t)])}{A(\theta)}\,\phi\,\theta\,f(t), \tag{5}$$

where $A'(\theta)=dA(\theta)/d\theta$ and $f(t)=-dS(t)/dt$. The hazard rate function (hrf) corresponding to (5) is

$$h_{\rm pop}(t)=\frac{A'(\theta[1-\phi F(t)])}{A(\theta[1-\phi F(t)])}\,\phi\,\theta\,f(t).$$

The model defined by Equations (3), (4) and (5) is called the *destructive power series cure rate* (DPSCR) model.

The (proper) survival function for the non-cured population represented by the rv T_{nc}, say $S_{\rm nc}(t)=P(T>t\,|N\geq 1)$, is

$$S_{\rm nc}(t)=\frac{A(\theta[1-\phi F(t)])-A(\theta[1-\phi])}{A(\theta)-A(\theta[1-\phi])},\ t>0. \tag{6}$$

Clearly, $S_{\rm nc}(0)=1$ and $S_{\rm nc}(\infty)=0$.

Equation (3) and the mixture cure rate model ([19], [1]) are related as

$$S_{pop}(t)=\frac{A(\theta[1-\phi])}{A(\theta)}+\left[1-\frac{A(\theta[1-\phi])}{A(\theta)}\right]S_{\rm nc}(t),$$

where $S_{\rm nc}(t)$ is given by (6). Thus, $S_{\rm pop}(t)$ is a mixture cure rate model with cure fraction p_0 and survival function $S_{\rm nc}(t)$.

Proposition 1. The survival function in (3) is identifiable.

Proof: Let $\boldsymbol{\vartheta_1} = (\phi_1, \theta_1, \boldsymbol{\lambda_1})$ and $\boldsymbol{\vartheta_2} = (\phi_2, \theta_2, \boldsymbol{\lambda_2})$ be such that $\boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta_2}$, where $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ are the parameters in $S(\cdot)$. If $S_{\text{pop}}(t|\boldsymbol{\vartheta_1}) = S_{\text{pop}}(t|\boldsymbol{\vartheta_2})$ for all $t > 0$, Equation (3) gives

$$\frac{A(\theta_1)}{A(\theta_2)} = \frac{A(\theta_1[1-\phi_1 F(t|\boldsymbol{\lambda}_1)])}{A(\theta_2[1-\phi_2 F(t|\boldsymbol{\lambda}_2)]))}, \ \forall t > 0. \qquad (7)$$

Since $A(\theta) = \sum_{m=1}^{\infty} a_m \theta^m$ is a monotone increasing function in θ and, without loss of generality, say $\theta_1 < \theta_2$, it follows that $A(\theta_1)/A(\theta_2) < 1$. For $\phi_1 \neq \phi_2$ and $\boldsymbol{\lambda_1} \neq \boldsymbol{\lambda_2}$, there exists t_0 such that $\theta_1[1-\phi_1 F(t_0|\boldsymbol{\lambda}_1)] > \theta_2[1-\phi_2 F(t_0|\boldsymbol{\lambda}_2)]$, and

$$A(\theta_1[1-\phi_1 F(t|\boldsymbol{\lambda}_1)])/A(\theta_2[1-\phi_2 F(t|\boldsymbol{\lambda}_2)]) > 1.$$

Therefore, the equalities in (7) do not hold, which completes the proof.

Some properties of the functions (3) and (6) are in the following remarks.

*Remark.*For any survival function $S(t)$, equation (3) can be reduced to

(i)$S_{\text{pop}}(t) = 1 - \phi + \phi S(t)$ when $\theta \to 0$,
(ii)$S_{\text{pop}}(t)$ is proper when $\phi \to 1$,
(iii)$S_{\text{pop}}(t) = S(t)$ when $\phi \to 1$ and $\theta \to 0$.

*Remark.*For any survival function $S(t)$, it follows from (6)

(i)$S_{\text{nc}}(t) = S(t)$ when $\theta \to 0$.

1.1 Some special models

Three special cases of the DPSCR model given by (3) are discussed here.

–If M is a rv having the zero truncated Poisson distribution, it follows from (3)

$$S_{\text{pop}}(t) = \frac{e^{-\theta\phi F(t)} - e^{-\theta}}{1 - e^{-\theta}}.$$

Then, the cure fraction is

$$p_0 = \lim_{t\to\infty} S_{\text{pop}}(t) = \frac{e^{-\theta\phi} - e^{-\theta}}{1 - e^{-\theta}}.$$

The corresponding pdf follows from (5) as

$$f_{\text{pop}}(t) = \frac{\theta\,\phi\, f(t) e^{-\theta\phi F(t)}}{1 - e^{-\theta}}.$$

–If M has a geometric distribution, it follows from (3)

$$S_{\text{pop}}(t) = \frac{(1-\theta)[1-\phi F(t)]}{1-[1-\phi F(t)]\theta}.$$

The cure fraction is $p_0 = (1-\phi)(1-\theta)[1-\theta(1-\phi)]^{-1}$ and the corresponding pdf becomes

$$f_{\text{pop}}(t) = (1-\theta)\,\phi\, f(t)\{1-\theta[1-\phi F(t)]\}^{-2}.$$

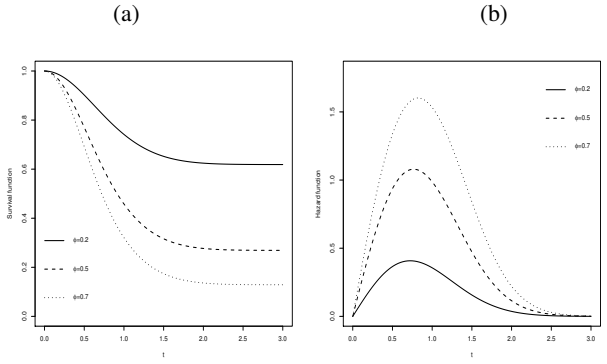

Fig. 2: Plots of the destructive Poisson ($\theta = 2$). (a) Survival functions. (b) Hazard functions.

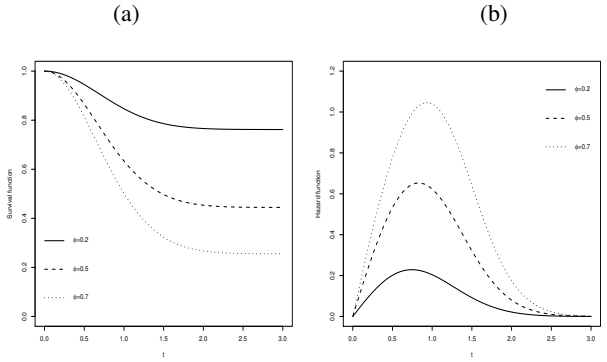

Fig. 3: Plots of the destructive geometric ($\theta = 0.2$). (a) Survival functions. (b) Hazard functions.

–If M has the logarithmic distribution,

$$S_{\text{pop}}(t) = \frac{\log\{1-[1-\phi F(t)]\,\theta\}}{\log(1-\theta)}.$$

The cure fraction is $p_0 = \log[1-(1-\phi)\theta]/\log(1-\theta)$ and the associated pdf is

$$f_{\text{pop}}(t) = -\frac{\phi\,\theta\, f(t)}{\log(1-\theta)\{1-\theta[1-\phi F(t)]\}}.$$

The plots displayed in Figures 2-4 (left panel) show different behaviors of the survival functions for the models defined before. Figures 2-4 (right panel) also reveal different shapes of the hazard rates for these models with $F(t) = 1 - e^{-t^2}$. These plots illustrate the flexibility afforded by the proposed model.

2 The DPSCR model for the non-cured population

The DPSCR model contains, as special cases, some known distributions. Several new ones can be easily generated.

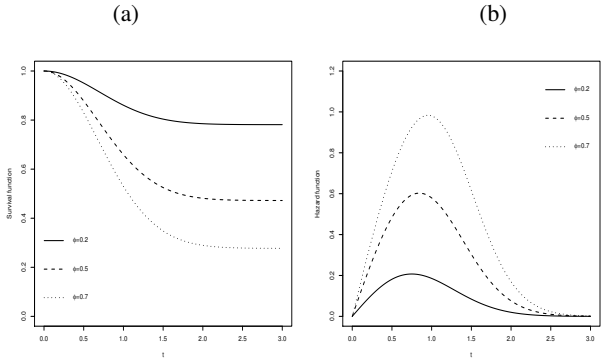

Fig. 4: Plots of the destructive logarithmic ($\theta = 0.2$). (a) Survival functions. (b) Hazard functions.

Equation (6) gives the proper survival function for the non-cured population and the density of T_{nc} has the form

$$f_{\rm nc}(t) = \frac{\theta\,\phi\,f(t)A'(\theta[1-\phi\,F(t)])}{A(\theta)-A(\theta[1-\phi])}. \qquad (8)$$

2.1 Linear representation

Using the power series for $A(\theta)$ given in Section 2, $A'(\theta) = \sum_{i=1}^{\infty} b_i\,\theta^i$, where $b_i = (i+1)\,a_{i+1}$ for $i \geq 0$. Then,

$$A'(\theta[1-\phi\,F(t)]) = \sum_{i=0}^{\infty} b_i\,\theta^i \sum_{j=0}^{i} \binom{i}{j} (-\phi)^j\,F(t)^j.$$

It can be written changing the order of the sums $A'(\theta[1-\phi\,F(t)]) = \sum_{j=0}^{\infty} u_j\,F(t)^j$, where $u_j = (-\phi)^j \sum_{i=j}^{\infty} \binom{i}{j} b_i\,\theta^i$. Moreover, it follows from (8)

$$f_{\rm nc}(t) = \sum_{j=0}^{\infty} s_j\,h_{j+1}(t), \qquad (9)$$

where $h_{j+1}(t) = (j+1)\,F(t)^j\,f(t)$ is the *exponentiated-F* (exp-F) density (with positive support) and power parameter $(j+1)$ (for $j \geq 0$) and

$$s_j = \frac{\theta\,\phi\,u_j}{(j+1)[A(\theta)-A(\theta[1-\phi])]}.$$

Hence, the density of T_{nc} depends on the power series distribution only through the coefficients s_j's.

The density function of T_{nc} in (9) is a linear combination of the exp-F densities and its properties can be obtained from those of V_{j+1} having density $h_{j+1}(t)$, which have been explored in many papers for several F baselines.

2.2 Moments and generating function

First, the nth raw moment of T_{nc} follows from (9) and the monotone convergence theorem. For $n \in \mathbb{N}$,

$$\mathrm{E}(T_{nc}^n) = \sum_{j=1}^{\infty} s_j\,\mathrm{E}(V_{j+1}^n).$$

The moments of V_{j+1} follow from the quantile function (qf) of the baseline F, say $Q_F(u) = F^{-1}(u)$, as $\mathrm{E}(V_{j+1}^n) = (j+1)\int_0^1 Q_F(u)^n\,u^j du$.

The nth incomplete moment of T_{nc} has the form

$$m_n(y) = \int_0^y t^n\,f_{\rm nc}(t)dt = \sum_{j=1}^{\infty}(j+1)\,s_j \int_0^{F(y)} Q_F(u)^n\,u^j du.$$

The generating function of T_{nc}, say $M(s)$, follows from (9) and the monotone convergence theorem as

$$M(s) = \sum_{j=0}^{\infty} s_j\,M_{j+1}(s),$$

where $M_{j+1}(s)$ is the generating function of V_{j+1}, namely $M_{j+1}(s) = (j+1)\int_0^{\infty} \mathrm{e}^{st}\,F(t)^j\,f(t)dt = (j+1)\int_0^1 \exp[s\,Q_F(u)]\,u^j du$.

3 Inference

Consider n cancer patients and let M_i be the number of carcinogenic cells before treatment for the i-th patient $(i = 1,\ldots,n)$. The M_i's are assumed to be iid rvs with pmf (1). Given $M_i = m_i$, let W_{ij} be independent Bernoulli rvs (independent of M_i for $j = 1,\ldots,m_i$) with success probability ϕ, indicating the presence of the j-th lesion (or undestroyed clonogenic cells). Let $N_i = W_{i1} + \cdots + W_{im_i}$ be the total number of altered cells among the M initial cells (competing causes) not destroyed by the treatment.

After the tumor is removed in surgery or another treatment, some carcinogenic cells can stay inside the body. If one of these cells is activated again, then cancer returns. Thus, we use the term "promotion time" to refer to the minimum time for one of the cells to become active. Furthermore, suppose that $Z_{i1},\cdots,Z_{iN_i}$ are the iid unobserved promotion times of the N_i carcinogenic cells for the i-th subject having proper cumulative distribution $F(\cdot|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of parameters. Let $y_i = \min\{T_i, C_i\}$ be the observed time, where $T_i = \min\{Z_{i1},\cdots,Z_{iNi}\}$, C_i is the censoring time, and δ_i is the censoring indicator equal to one if $y_i = T_i$ and zero otherwise (a right-censored time).

Consider the systematic component $\phi_i = \exp(\boldsymbol{x}_{1i}^{\top}\boldsymbol{\beta})/[1+\exp(\boldsymbol{x}_{1i}^{\top}\boldsymbol{\beta})]$, where $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector to estimate the effects of the covariates on the proportion of undestroyed clonogenic cells. Also, based on the proportional hazard function, the covariate vector $\boldsymbol{x}_{2i}$ can be incorporated

$$h(y_i|\boldsymbol{\lambda}) = h_0(y|\alpha)\,\exp(\boldsymbol{x}_{2i}^{\top}\boldsymbol{\gamma}), \qquad (10)$$

where $\boldsymbol{\lambda}^\top = (\alpha, \boldsymbol{\gamma}^\top)$, $h_0(y|\alpha)$ is the baseline hazard function, α is a positive scale parameter of the baseline function and $\boldsymbol{\gamma}$ is a $r \times 1$ vector of unknown coefficients. Considering the baseline hazard function $h_0(y|\alpha) = \alpha y^{\alpha-1}$, the Z_{ij}'s have a Weibull density $f(y_i|\boldsymbol{\lambda}) = \alpha y_i^{\alpha-1} \exp[\boldsymbol{x}_{2i}^\top \gamma - y_i^\alpha \exp(\boldsymbol{x}_{2i}^\top \gamma)]$ with cdf $F(y_i|\boldsymbol{\lambda}) = 1 - \exp\left[-\alpha y_i^\alpha \exp(\boldsymbol{x}_{2i}^\top \gamma)\right]$.

Then, the likelihood function for $\boldsymbol{\vartheta} = (\theta, \alpha, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$, under the Weibull distribution, can be reduced to

$$L(\boldsymbol{\vartheta}|\mathscr{D}) = \prod_{i=1}^{n} \left\{ \frac{\theta A'(\theta[1-\phi_i F(y_i|\alpha,\boldsymbol{\gamma})])\phi_i f(y_i|\alpha,\boldsymbol{\gamma})}{A(\theta[1-\phi_i F(y_i|\alpha,\boldsymbol{\gamma})])} \right\}^{\delta_i} \times \left\{ \frac{A(\theta(1-\phi_i)F(y_i|\alpha,\boldsymbol{\gamma}))}{A(\theta)} \right\}, \quad (11)$$

where $\mathscr{D} = (\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{x}_1, \boldsymbol{x}_2)$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{x}_1 = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n})^\top$, $\boldsymbol{x}_2 = (\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n})^\top$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$.

The log-likelihood function is $\ell(\boldsymbol{\vartheta}) = \log[L(\boldsymbol{\vartheta}|\mathscr{D})]$ and the components of the score vector $U(\boldsymbol{\vartheta})$ can be available from the authors on request. The MLE $\widehat{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}$ can be found by solving the nonlinear equation system $U_\theta(\boldsymbol{\vartheta}) = 0$, $U_\alpha(\boldsymbol{\vartheta}) = 0$, $U_{\beta_j}(\boldsymbol{\vartheta}) = 0$ and $U_{\gamma_{j'}}(\boldsymbol{\vartheta}) = 0$ in the statistical software R, whose script is also available from the authors.

4 Simulation study

We perform a simulation study to examine the precision of the MLEs in the DPSCR model (3). The number of causes (M_i) not destroyed for the i-th patient is generated from a truncated Poisson distribution with $\theta = 2$ (for $i = 1, \ldots, n$). Given $M_i = m_i$, the Bernoulli rv W_{ij} ($j = 1, \ldots, m_i$) is generated with success probability $\phi_i = \exp(\beta_0 + \beta_1 x_i)/[1 + \exp(\beta_0 + \beta_1 x_i)]$, where $\beta_0 = -1.0$ and $\beta_1 = 1.5$. The total number of destroyed causes among the m_i initial causes is $N_i = W_{i1} + \cdots + W_{im_i}$. Given $N_i = n_i$, the event times Z_{ij} (for $j = 1, \ldots, n_i$) are generated from a Weibull hazard function $h(z_{ij}|\alpha, \boldsymbol{\gamma}) = \alpha z_{ij}^{\alpha-1} \exp(\gamma_0 + \gamma_1 x_i)$, where $\alpha = 2.0$, $\gamma_0 = 2.0$ and $\gamma_1 = 1.5$. The censoring times are sampled from the Uniform$(0, \tau)$ distribution, where τ is approximately equal to 56%.

One thousand simulations are run for each sample size $n = 200, 400, 600$ and $1,000$ to calculate the average of the MLEs (AMLE), standard error (SE), bias, root of mean squared error (RMSE) and empirical coverage probability (CP) corresponding to the nominal 95% confidence interval for the parameters. The results are reported in Table 1. The averages of the estimates converge to the true parameter values, the RMSEs and biases decrease and the coverage probabilities become much closer to the nominal level when n increases, which show that the estimates are consistent and approximately normal.

Table 1: Simulation results from the DPSCR model in (3).

n		θ	α	γ_0	γ_1	β_0	β_1
200	AMLE	1.412	1.979	1.090	1.988	-0.731	1.672
	SE	1.282	0.167	0.292	0.430	0.531	0.38
	BIAS	-0.588	-0.021	0.090	-0.012	0.269	0.172
	RMSE	1.410	0.168	0.305	0.430	0.595	0.420
	CP	0.812	0.950	0.931	0.940	0.831	0.961
400	AMLE	1.632	1.982	1.061	1.988	-0.817	1.592
	SE	1.180	0.121	0.205	0.290	0.469	0.279
	BIAS	-0.368	-0.018	0.061	-0.012	0.183	0.092
	RMSE	1.235	0.122	0.214	0.291	0.504	0.293
	CP	0.902	0.947	0.939	0.941	0.912	0.960
600	AMLE	1.767	1.986	1.043	1.991	-0.875	1.556
	SE	1.074	0.100	0.158	0.235	0.413	0.241
	BIAS	-0.233	-0.014	0.043	-0.009	0.125	0.056
	RMSE	1.098	0.101	0.163	0.236	0.431	0.247
	CP	0.935	0.957	0.956	0.954	0.955	0.949
1000	AMLE	1.936	1.991	1.024	1.992	-0.949	1.526
	SE	0.992	0.075	0.125	0.191	0.370	0.187
	BIAS	-0.064	-0.009	0.024	-0.008	0.051	0.026
	RMSE	0.993	0.076	0.127	0.191	0.373	0.189
	CP	0.936	0.958	0.968	0.948	0.960	0.957

5 Application to real data

Consider the survival time (T) until the patient's death (in years) or the censoring time at the end of the study, as already discussed in Section 1, for application of the new regression. The data set refers to $n = 365$ women who underwent mastectomy mastectomy at Memorial Sloan-Kettering Cancer Center (New York) between 1976 and 1979 [15]. The immunohistochemistry (IHC) and hematoxylin and eosin (H&E) stains measure the lymph node status and the following variables were obtained: y_i: observed time (in years); x_{i1}: age (in years); x_{i2}: multifocality (0: no, 1:yes); x_{i3}: tumor size (in cm); x_{i4}: tumor grading (0: I, 1: II, II and lobular); x_{i5}: lymphovascular invasion (0: no, 1: yes) and x_{i6}: lymph node status (0: IHC+ IHC- and H&E-, 1: IHC+ and H&E+).

It is taken $x_{i1} = x_{1i1} = x_{2i1}$, $x_{i2} = x_{1i2} = x_{2i2} = \cdots = x_{1i6} = x_{2i6}$ and the systematic components are expressed as (for $i = 1, \ldots, 284$)

$$\phi_i = \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}\}}$$

and

$$h(y_i|\alpha,\gamma) = \alpha y_i^{\alpha-1} \times \exp\{\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i6}\}.$$

We consider a reduced sample of $n = 284$ patients (78% of censoring) after deleting patients with incomplete data and missing observation times. The Kaplan–Meier estimate in Figure 1 (left panel) by tumor grading has a well pronounced plateau at the level above zero according to [2].

Different regressions are compared via the Akaike information criterion AIC $= -2\ell(\widehat{\boldsymbol{\vartheta}}) + 2\#(\boldsymbol{\vartheta})$ and Schwartz–Bayesian criterion

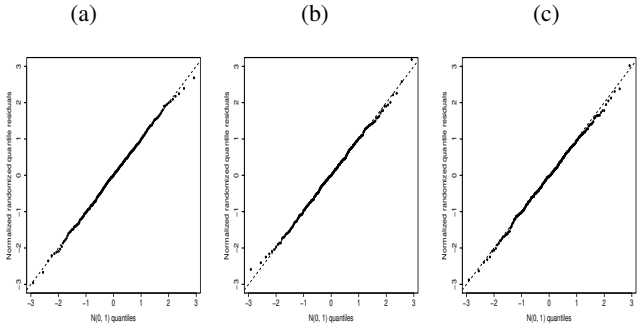

Fig. 5: QQ plots.

$\text{SBC} = -2\ell(\widehat{\boldsymbol{\vartheta}}) + \#(\boldsymbol{\vartheta})\log(n)$, where $\#(\boldsymbol{\vartheta})$ is the number of estimated parameters. The special models discussed in Section 1 are fitted to these data.

The selection criteria for three fitted regressions are given in Table 2. Each point in the QQ plots of the quantile residuals ([20], [21]) for the destructive Poisson, geometric and logarithmic models with identity link function in Figure 5 corresponds to the median of five sets of ordered residuals. The statistics in Table 2 and these QQ plots indicate that the destructive geometric cure rate (DGCR) regression is the best model to fit these data.

Table 2: Selected criteria for the fitted models.

Model	$\ell(\widehat{\boldsymbol{\vartheta}})$	AIC	SBC
Poisson	-308.314	648.627	707.011
Geometric	-307.760	647.520	705.903
Logarithmic	-308.154	648.307	706.691

The MLEs for the full model are reported in Table 3. The estimate of α rejects the exponential distribution ($\alpha = 1$) for the unobserved failure times.

The effects of the covariates in the probability of undestroyed causes and short-term survivors can be based on the likelihood ratio statistic for testing $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \beta_1 = \beta_2 = 0$, which yields $w_n = 0.179$ (p-value=1). Thus, the age, multifocality, tumor size, tumor grading, lymphovascular invasion and lymph node status are insignificant for the probability of undestroyed causes. Similarly, the age and multifocality do not have significant effects on short-term survivors. Thus, the MLEs (and their SEs) of the parameters for the reduced DGCR regression with significant covariates are reported in Table 4, where the values of AIC and SBC are 631.699 and 660.891, respectively. Comparing these numbers with the figures in Table 2, we conclude that the reduced DGCR regression provides a similar fit to these data.

Table 3: Results from the fitted DGCR regression.

Parameter	MLE	SE	|MLE| / SE
θ	0.134	0.335	–
α	2.043	0.225	–
γ_0	-2.926	1.781	1.643
γ_1	0.006	0.019	0.333
γ_2	-0.636	0.543	1.172
γ_3	0.133	0.238	0.557
γ_4	-1.712	1.730	0.990
γ_5	-0.666	0.617	1.080
γ_6	-0.582	0.444	1.313
β_0	-2.854	1.067	2.674
β_1	-0.017	0.014	1.209
β_2	0.764	0.527	1.450
β_3	0.274	0.148	1.850
β_4	1.839	0.684	2.688
β_5	0.432	0.479	0.903
β_6	1.665	0.550	3.026

Table 4: MLEs of the parameters for the reduced DGCR regression.

Parameter	Estimate (est)	Standard error (se)	|est| / se
θ	0.005	0.076	—
α	1.933	0.205	—
γ_0	-4.075	0.463	8.805
β_0	-5.216	1.163	4.483
β_3	0.391	0.168	2.331
β_4	3.165	1.102	2.871
β_5	0.719	0.394	1.824
β_6	1.900	0.518	3.666

The DGCR regression is now fitted by considering just one covariate. The empirical and estimated survival functions for each covariate are displayed in Figures 6 (a), (b), (c). In fact, the reduced DGCR regression provides a good fit to these data.

The following calculations are for illustrative purposes. The explanatory variables for four hypothetical mastectomized women A, B, C and D are given in Table 5. For example, because of these variables there are different cure rates and probability of undestroyed causes, namely 0.979 and 0.021 for woman A and 0.227 and 0.772 for woman D. The right and center panels of Figure 7 display plots of the estimated probabilities of undestroyed causes and cure rates for four mastectomized women in terms of characteristics of the tumors by fixing the size. The plots reveal that the probability of undestroyed causes increases with increasing size tumor and cure rate decreases more rapidly with increasing tumor size. The plots in Figure 7 display some estimated functions for these women described above.

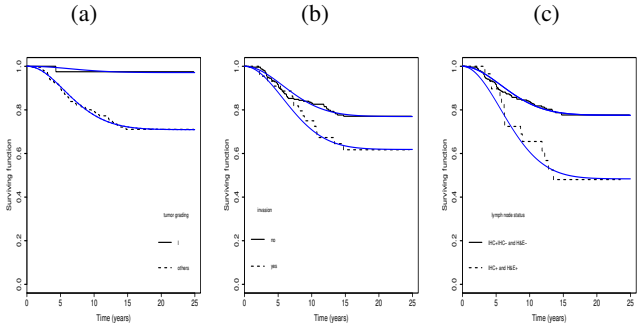

Fig. 6: Kaplan-Meier curves and estimated survival functions stratified by: (a) tumor grading. (b) lymphovascular invasion. (c) lymph node status.

Table 5: Estimated probability of undestroyed causes and cure rates for four mastectomized women.

Patient	tumor size	tumor grade	lymphovascular	staining	p_0	ϕ
A	3.5	0	0	0	0.979	0.021
B	3.5	0	0	1	0.875	0.125
C	3.5	0	1	1	0.772	0.227
D	3.5	1	0	1	0.227	0.772

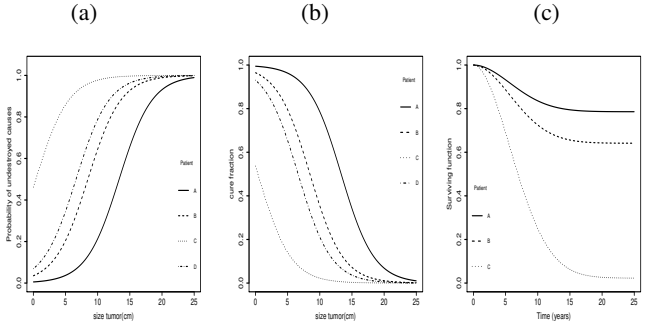

Fig. 7: Probability of undestroyed causes (right panel), cure rates (center) and surviving function (right panel) under the DGCR model for four mastectomized women.

6 Concluding remarks

We proposed a destructive power series cure rate model which allows to estimate the proportion of causes not eliminated by initial treatment (undestroyed causes). Some compounding regressions are special cases of the introduced formulation, namely: the destructive Poisson, logarithmic and geometric cure rate models. The recurrence time distribution for the entire population was investigated and some mathematical properties of the recurrence time for the non-cured individuals are addressed. The regression model was appropriate to study the recurrence time and cure fraction of several types of cancers after surgery. An application was provided to evaluate the risk of breast cancer recurrence after the mastectomy by assuming that the promotion times of the carcinogenic cells followed the Weibull distribution. The maximum likelihood estimation method provided consistent estimators of the regression coefficients. We showed that the tumor size, tumor grading, lymphovascular invasion and lymph node status are significant prognostic variables for survival time and mortality risk in mastectomized women.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] J. Berkson, and R.P. Gage, *Survival curve for cancer patients following treatment*, *Journal of the American Statistical Association*. **259**, 501-515 (1952).

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

[3] J.G. Ibrahim, M.H. Chen, and D. Sinha, *Bayesian Survival Analysis*, Springer, New York, (2001).

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

[5] J. Rodrigues, M. de Castro, N. Balakrishnan, and V.G. Cancho, Destructive weighted Poisson cure rate models, *Lifetime Data Analysis*. **17**, 333-346 (2010).

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

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

[8] J.B. Fachini, E.M.M. Ortega, and G.M. Cordeiro, A bivariate regression model with cure fraction, *Journal of Statistical Computation and Simulation.* **84**, 1580-1595 (2014).

[9] E.M.M. Ortega, G.M. Cordeiro, A.K. Campelo, M.W. Kattan, and V.G. Cancho, A power series beta Weibull regression model for predicting breast carcinoma, *Statistics in Medicine.* **34**, 1366-1388 (2015).

[10] E.M. Hashimoto, E.M.M. Ortega, V.G. Cancho, and G.M. Cordeiro, A new long-term survival model with interval-censored data, *Sankhya B*. **77**, 207-239 (2015).

[11] B. Yiqi, V.G. Cancho, and F. Louzada-Neto, On the Bayesian estimation and influence diagnostics for the Weibull-Negative-Binomial regression model with cure rate under latent failure causes, *Communications in Statistics - Theory and Methods*. **46**, 1462-1489 (2017).

[12] T.G. Ramires, N. Hens, G.M. Cordeiro, and E.M.M. Ortega, Estimating nonlinear effects in the presence of cure fraction using a semi-parametric regression model, *Computational Statistics*. **33**, 709-730 (2017).

[13] E.M.M. Ortega, G.M. Cordeiro, E.M. Hashimoto, and A.K. Suzuki, Regression models generated by gamma random variables with long-term survivors, *Communications for Statistical Applications and Methods*. **24**, 43-65 (2017).

[14] G.D.C. Barriga, V.G. Cancho, D.V. Graibay, G.M. Cordeiro, and E.M.M. Ortega, A new survival model with surviving fraction: An application to colorectal cancer data, *Statistical Methods in Medical Research*. **28**, 2665-2680 (2018).
[15] M.W. Kattan, D. Giri, K.S. Panageas, A. Hummer, M. Cranor, K.J. Van Zee, C.A. Hudis, L. Norton, P.I. Borgen, and L.K. Tan, A tool for predicting breast carcinoma mortality in women who do not receive adjuvant therapy, *Cancer*. **101**, 2509-2515 (2004).
[16] L.N. Johnson, A.W. Kemp, and S. Kotz, *Univariate Discrete Distribution*, Wiley, New York, NY, third edition, (2005).
[17] G. Yang, and C. Chen, A stochastic two-stage carcinogenesis model: a new approach to computing the probability of observing tumor in animal bioassays, *Mathematics Biosience*. **104**, 247-258 (1991).
[18] A.D. Tsodikov, J.G. Ibrahim, and A.Y. Yakovlev, Estimating cure rates from survival data: An alternative to two-component mixture models, *Journal of the American Statistical Association*. **98**, 1063-1078 (2003).
[19] J.W. Boag, Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society B*. **11**, 15-53 (1949).
[20] P.K. Dunn and G.K. Smyth, Randomized quantile residuals, *Journal of Computational and Graphical Statistics*. **5**, 236-244 (1996).
[21] R.A. Rigby, and D.M. Stasinopoulos, Generalized additive models for location, scale and shape (with discussion), *Applied Statistics*. **54**, 507-554 (2005).

Vicente G. Cancho Vicente G. Cancho is PhD in Statistics from University of São Paulo (1999). He is currently professor at the Department of Applied Mathematics and Statistics Institute of Mathematics and Computer Sciences University of São Paulo at Sao Carlos, Brazil. He works on Bayesian methods, regression models and survival analysis..

Gauss M. Cordeiro is PhD in Statistics from Imperial College, University of London (1982). He has a postdoctoral degree from the University of London (1986-1987) and the Institute of Pure and Applied Mathematics (1990-1992). He is currently a full professor at the Department of Statistics at the Federal University of Pernambuco. He has published 433 articles in international statistics journals with peer review systems and has supervised 59 master and doctoral students. He works on distribution theory and asymptotic theory. He has worked as a reviewer for several scientific journals since 1983. He was President of the Brazilian Statistical Association (ABE) from 2000 to 2002 and member of the Mathematical Advisory Committee of the National Council for Scientific and Technological Development in four different periods. He was editor of the Brazilian Journal of Probability and Statistics (1995-2000) and received the ABE-2008 Prize for outstanding services to Statistics and, in 2010, the degree of Commander of the Order of Scientific Merit.

Gladys Barriga Gladys D.C. Barriga is PhD in Production Engineering .from University of São Paulo (2001). she is currently a professor at the Department of Production Engineering in São Paulo State University at Bauru, Brazil. He works on industrial statistics, regression models, and realibility analysis.

Edwin M. M. Ortega is PhD in Statistics from, University of São Paulo (2002). He has a postdoctoral degree from the Federal University of Pernambuco (2010-2011). He is currently a full professor at the Department of Exact Sciences at the University of São Paulo, ESALQ-USP, Brazil. He has published 205 papers in international statistics journals with peer review systems and has supervised master and doctoral students. He works on regression models, semiparametric regression models, distribution theory and survival analysis. He has worked as a reviewer for several scientific journals since 2005.

Michael W. Kattan Michael W. Kattan, Ph.D. is Chairman of the Department of Quantitative Health Sciences at The Cleveland Clinic and Professor of Medicine, Epidemiology and Biostatistics, Cleveland Clinic Lerner College of Medicine of Case Western Reserve University. He is also the Dr. Keyhan and Dr. Jafar Mobasseri Endowed Chair for Innovations in Cancer Research.