Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data

Patrick Borges[a], Josemar Rodrigues[b] and Narayanaswamy Balakrishnan[c]

[a]Departamento de Estatística, Universidade Federal do Espírito Santo, Brazil
[b]Departamento de Estatística, Universidade Federal de São Carlos, Brazil
[c]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada

Abstract

In this paper, we propose a new cure rate survival model, which extends the model of Rodrigues et al. (2011) by incorporating a dependence structure between the initiated cells. To create the correlation structure between the initiated cells, we use an extension of the generalized power series distribution by including an additional parameter ρ (inflated-parameter generalized power series (IGPS) distribution, studied by Kolev (2000)). It has a natural interpretation in terms of both "zero-inflated" proportion and correlation coefficient. In our approach, the number of initiated cells is assumed to follow the IGPS distribution. The IGPS distribution is a natural choice for modeling correlated count data that exhibit overdispersion. The primary advantage of this distributional assumption is that the correlation structure induced by the additional parameter ρ results in a natural characterization of the association between the initiated cells. Moreover, it provides a simple and realistic interpretation for the biological mechanism of the occurrence of the event of interest as it includes a destructive process of tumor cells after an initial treatment or the capacity of an individual exposed to irradiation to repair initiated cells that results in cancer induction. This means that, what is recorded is only the undamaged portion of the original number of initiated cells not eliminated by the treatment or repaired by the repair system of an individual. Parameter estimation of the proposed model is then discussed through the maximum likelihood estimation procedure. Finally, we illustrate the usefulness of the proposed model by applying it to a real cutaneous melanoma data.

Key words: Initiated cells, Cure rate models, IGPS distribution, Additional parameter ρ, Correlation structure.

1. Introduction

Cure rate survival modelling plays an important role in reliability and survival analysis. It pertains to survival studies wherein a proportion of the subjects might not be susceptible to the event of interest due to different competing causes. These models have found important applications in such diverse

*Corresponding author: Patrick Borges. Departamento de Estatística, Universidade Federal do Espírito Santo, Av. Fernando Ferrari 514, Goiabeiras, CEP 29075-910, Vitória, Espírito Santo, Brasil. Email: patrick@cce.ufes.br

areas as biomedical studies, finance, criminology, demography, manufacturing and industrial reliability. For instance, in biomedical data, an event of interest can be a patient's death, which can occur due to different competing causes or a tumor recurrence that may arise due to a number of metastasis-component tumor cells that are left active after an initial treatment of the individual. A metastasis-component tumor cell is a tumor cell which has the potential of metastasizing; see Yakovlev (1994), Yakovlev et al. (1993), Yakovlev and Tsodikov (1996), and Ibrahim et al. (2001).

The classical Berkson-Gage model (Berkson and Gage, 1952), discussed further by Farewell (1982, 1986), Goldman (1984), Sy and Taylor (2000), and Banerjee and Carlin (2004), as well as the most recent and comprehensive models of Yakovlev and Tsodikov (1996), Chen et al. (1999), Ibrahim et al. (2001), Chen et al. (2002), and Yin and Ibrahim (2005) include in their formulation the possibility of evaluating the cured rate. It is reasonable to presume that the occurrence of the event of interest might be due to one of many competing causes (Gordon, 1990), with the number of causes and the distribution of survival times associated with each cause (Cox and Oakes, 1984, p.147) being unknown, which leads to the so-called latent competing causes. Another approach, by Cooner et al. (2007), stochastically forms an arranged sequence of latent causes that induces the occurrence of the event of interest.

Cure rate models, with the number of latent competing causes following a Poisson, negative binomial, geometric or a Bernoulli distribution, have been studied by Rodrigues et al. (2009), Chen et al. (1999), and Cooner et al. (2007). The well-known Berkson-Gage model (Berkson and Gage, 1952) corresponds to case when the number of latent causes follows a Bernoulli distribution and that there is at most one latent cause.

However, in most survival models incorporating surviving fraction that are used in analyzing data from cancer clinical trials, there are two basic weaknesses:

1. first, the assumption that each initiated cell (competing cause or risk factor) becomes cancerous with probability one, and
2. next, the assumption of biological independence of initiated cells while becoming cancerous.

For overcoming Weakness 1 mentioned above, in a recent paper, Rodrigues et al. (2011), motivated by the work of Klebanov et al. (1993), proposed a stochastic damaged model for survival data with a surving fraction (also known as destructive weighted Poisson cure rate models) for describing the biological process of elimination of initiated cells after some specific treatment, but assuming biological independence of cells. With regard to Weakness 2, Haynatzki et al. (2000) pointed out that the biological independence assumption may not be valid when the dynamics of the cell population of a normal tissue is considered. Similarly, there are also indications that premalignant (initiated) and malignant cells in a tissue influence each other's development to some extent. Moreover, the interaction between healthy and premalignant cells in the tissue should not be excluded from consideration either. It is, therefore, quite desirable to construct mathematically tractable models that can adequately incorporate biological dependence, and this is the primary motivation for the present research work.

The main purpose of this paper is to propose a new cure survival model which extend the models of Rodrigues et al. (2011) by incorporating a dependence structure between the initiated cells. To create the dependence structure between these initiated cells, we use an extension of the generalized power series distributions by including an additional parameter ρ (inflated-parameter generalized power

series (IGPS) distribution; see Kolev (2000)). This has a natural interpretation in terms of both "zero-inflated" proportion and correlation coefficient. In our approach, we assume the number of initiated cells to follow an IGPS distribution, which is a natural choice for modeling correlated count data that possess overdispersion. The advantage of this assumption is that the correlation structure induced by this additional parameter ρ of the model results in a natural characterization of the association between the initiated cells. Moreover, it provides a simple and meaningful interpretation of the underlying biological mechanism of the occurrence of the event of interest as it includes a destructive process of tumor cells after an initial treatment or the capacity of an individual exposed to irradiation to repair initiated cells that results in cancer induction. In other words, what is recorded is only the undamaged portion of the original number of initiated cells not eliminated by the treatment or repaired by the repair system of an individual, which is represented by a compound variable.

The rest of the paper is organized as follows. In Section 2, we describe the model formulation. Some specific models and special cases are given in Section 3. In Section 4, we discuss the maximum likelihood estimation of the model parameters. In Section 5, a data on melanoma is used to illustrate the usefulness of the proposed model. Finally, some concluding comments are made in Section 6.

2. Model formulation

For an individual in the population, let N denote the number of initiated cells related to the occurrence of a tumor. Assume that the unobserved latent variable N has an inflated-parameter generalized power series (IGPS) distribution with probability mass function $(p.m.f.)$ given by

$$p(n;\theta,\rho) = \mathbb{P}[N=n;\theta,\rho] = \frac{1}{g(\theta)} \sum_{n_1,n_2,\ldots} a_n \big[\theta(1-\rho)\big]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \quad n=0,1,2,\ldots,\ \rho\in[0,1), \quad (1)$$

where a_n depends only on n, $g(\theta) = \sum_{n=0}^{\infty} a_n\theta^n$ is a positive, finite and differentiable function and $\theta \in (0,s)$ (s can be ∞) is such that $g(\theta)$ is finite, and the summation is over the set of all nonnegative integers $n_1, n_2, \ldots$ such that $\sum_{i=1}^{\infty} in_i = n$. For more details on the IGPS distribution, one may refer to Kolev (2000) and Minkova (2002). The parameter ρ is a measure of association between the tumor cells. Large values of ρ indicate high association between cells, while $\rho \to 0$ implies less association between cells. It is of interest to note that when $\rho = 0$ (i.e., when there is independence between cells), the IGPS distribution reduces to a generalized power series distribuion (Gupta, 1974; Consul, 1990). Table 5.1 presents the choices of a_n, $g(\theta)$ and the parameter θ corresponding to some special cases of IGPS distributions, namely, inflated-parameter Poisson (IP), negative binomial (INB), binomial (IB) and logarithmic-series (ILS) distributions. In the IB and INB cases, the additional parameters $m \in \mathbb{Z}^+$ (set of non-negative integers) and $\phi > -1$ are to be treated as nuisance parameters.

Table 2.1: The choices of a_n, $g(\theta)$ and the parameter θ for some special cases of IGPS distributions.

Distribution	a_n	$g(\theta)$	θ	s
IP	$\frac{1}{n_1!n_2!\cdots}$	e^{θ}	η	∞
IB	$\binom{m}{m-n_1-n_2-\cdots,n_1,n_2,\ldots}$	$(1+\theta)^m$	$\frac{\pi}{1-\pi}$	1
INB	$\frac{\Gamma\left(\phi^{-1}+\sum_{i=1}^{\infty} n_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} n_i\right]!}$	$(1-\theta)^{-\phi^{-1}}$	$\frac{\phi\eta}{1+\phi\eta}$	∞
ILS	$\frac{(-1+n_1+n_2+\cdots)!}{n_1!n_2!\cdots}$	$-\log(1-\theta)$	$1-\pi$	1

The probability generating function ($p.g.f.$) of the inflated-parameter generalized power series random variable N is given by

$$\mathbb{A}_N(z) = \frac{g\left(\theta z(1-\rho)(1-z\rho)^{-1}\right)}{g(\theta)} \quad \text{for} \quad 0 \leq z \leq 1. \tag{2}$$

Now, after a prolonged treatment, we have as an immediate consequence the formation of possibly precancerous lesions in a genome of the cells. Let $N = n$ be the number of such lesions or initiated cells after the treatment, and X_j, $j = 1, 2, \ldots, n$, be independent random variables, independently of N, having a Bernoulli distribution with success probability p indicating the presence of the j^{th} lesion with $p.g.f.$

$$\mathbb{A}_{X_j}(z) = 1 - p(1-z), \quad \text{for} \quad 0 \leq z \leq 1. \tag{3}$$

The variable D, representing the total number of cells, among the N initiated cells, not eliminated by the treatment, is then given by

$$D = \begin{cases} \sum_{j=1}^{N} X_j & , \quad if \quad N > 0 \\ 0 & , \quad if \quad N = 0 \end{cases}. \tag{4}$$

By damaged or unrepaired irradiation, we mean that $D \leq N$.

This viewing of (4) has been suggested earlier by Yang and Chen (1991) in the context of a biossay study. They assumed that the initial risk factors are primary initiated malignant cells, where X_j in (4) denotes the number of living malignant cells that are descendants of the j^{th} initiated malignant cell during some time interval. In this context, D then denotes the total number of living malignant cells at some specific time.

In the competing causes scenario (Cox and Oakes, 1984), the number of unrepaired lesions D in (4) and the time V to transform theses lesions into a detectable tumor are both not observable (latent variables). We will call V a progression time. So, the time from the start of the treatment to tumor detection (which is the event of interest) in a given individual is defined by the random variable

$$Y = \min\{V_1, V_2, \ldots, V_D\} \tag{5}$$

for $D \geq 1$, and $Y = \infty$ if $D = 0$, which leads to a proportion p_0 of the population whose lesions are repaired by the treatment, also called the "cured fraction". We assume that $V_1, V_2, \ldots$ are independent of D, and that, conditional on D, the variables V_j are i.i.d..

According to Rodrigues et al. (2011), the long-term survival function of the random variable Y in (5) is given by

$$S_{pop}(y) = P[Y \geq y] = \mathbb{A}_D(S(y)) = \sum_{d=0}^{\infty} P[D = d]\{s(Y)\}^d = \mathbb{A}_N\Big(\mathbb{A}_{X_j}\big(S(y)\big)\Big),$$

where $S(\cdot)$ denotes the common survival function of the unobserved lifetimes in (5) and $\mathbb{A}_D(\cdot)$ is the probability generating function of the compound variable D, which converges when $z = S(y) \in [0, 1]$. Upon taking into account (2) and (3), the long-term survival function of the observed time of a detectable tumor in (5) is expressed by

$$S_{pop}(y) = \frac{g\Big(\theta\big(1-\rho\big)\big(1-pF(y)\big)\big[1-\big(1-pF(y)\big)\rho\big]^{-1}\Big)}{g(\theta)}, \tag{6}$$

where $F(y) = 1 - S(y)$. If we take specifically $\rho = 0$, we get the generalized power series long-term survival function.

Given a proper survival function $S(\cdot)$, we have

$$\lim_{y\to\infty} S_{pop}(y) = p_0 = \frac{g\big(\theta\big(1-\rho\big)\big(1-p\big)\big[1-\big(1-p\big)\rho\big]^{-1}\big)}{g(\theta)}, \tag{7}$$

where p_0 denotes the proportion of "cured" or "immune" individuals present in the population from which the sample data arose.

We refer to the model in (6) by the correlated destrutive generalized power series model, or simply the CDGPS model. Figure 2.1 describes the CDGPS model in a diagram matrix form:

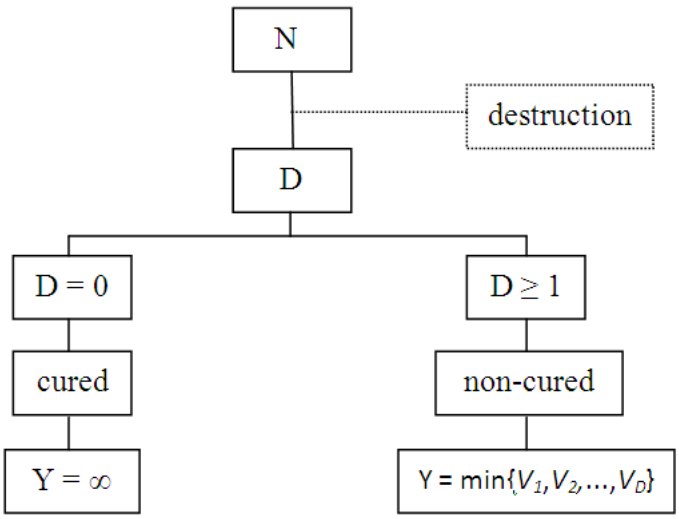

Figure 2.1: Representation of the proposed CDGPS model in a diagram matrix form.

3. Special cases of the proposed model

In this section, we present some special cases of the CDGPS model, proposed in the preceding section.

3.1. Correlated destrutive Poisson (CDP) model

For the choice of $a_n = \frac{1}{n_1!n_2!\cdots}$, $g(\theta) = \exp\{\theta\}$ and the parameter $\theta = \eta$, we say that the number of initiated cells N has an inflated-parameter Poisson distribution with parameters $\eta > 0$ and $\rho \in [0, 1)$, and its *p.m.f.* is of the form

$$\mathbb{P}_{Poi}[N = n] = \sum_{n_1,n_2,\cdots} \frac{e^{-\eta}}{n_1!n_2!\cdots}\Big[\eta(1-\rho)\Big]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{8}$$

where $n = 0, 1, 2, \ldots$, and the summation is over all nonnegative integers $n_1, n_2, n_3, \ldots$, such that $\sum_{i=1}^{\infty} in_i = n$. An alternate expression for the *p.m.f.* in (8) is given by

$$\mathbb{P}_{Poi}[N = n] = \begin{cases} e^{-\eta} & , \quad n = 0 \\ e^{-\eta}\sum_{i=1}^{n}\binom{n-1}{i-1}\frac{\left[\eta(1-\rho)\right]^i \rho^{n-1}}{i!} & , \quad n = 1, 2, \ldots. \end{cases} . \tag{9}$$

The mean and the variance of N are

$$\mathbb{E}[N] = \frac{\eta}{1-\rho} \quad \text{and} \quad \mathbb{V}ar[N] = \frac{\eta(1+\rho)}{(1-\rho)^2}, \tag{10}$$

respectively. The *p.g.f.* is given by

$$\mathbb{A}_N(z) = \exp\left\{-\frac{\eta(1-z)}{1-z\rho}\right\} \quad \text{for} \quad 0 \le z \le 1, \tag{11}$$

while the long-term survival function of the CDP model is given by

$$S_{pop}(y) = \exp\left\{-\frac{\eta p F(y)}{1-\rho(1-pF(y))}\right\}. \tag{12}$$

There are two important special cases of (12). For $\rho = 0$, we deduce the destructive Poisson model (Rodrigues et al., 2011), while for $\rho = 0$ and $p = 1$, we deduce the promotion time cure model (Yakovlev and Tsodikov, 1996; Chen et al., 1999).

3.2. Correlated destructive binomial (CDB) model

For the choice of $a_n = \binom{m}{m-n_1-n_2-\cdots,n_1,n_2,\ldots}$, $g(\theta) = (1+\theta)^m$ and $\theta = \frac{\pi}{1-\pi}$, the number of initiated cells N has an inflated-parameter binomial distribution with parameters $\pi \in (0, 1)$, $\rho \in [0, 1)$ and $m \in \mathbb{Z}^+$, and its *p.m.f.* is of the form

$$\mathbb{P}_{Bin}[N = n] = (1-\pi)^m \sum_{n_1,n_2,\ldots} \binom{m}{m-n_1-n_2-\cdots,n_1,n_2,\ldots} \rho^n \left\{\frac{\pi(1-\rho)}{\rho(1-\pi)}\right\}^{\sum_{i=1}^{\infty} n_i}, \tag{13}$$

where $n = 0, 1, \ldots$, and the summation is over all nonnegative integers $n_1, n_2, \ldots$, such that $\sum_{i=1}^{\infty} in_i = n$. An alternate expression for the *p.m.f.* in (13) is given by

$$\mathbb{P}_{Bin}[N = n] = \begin{cases} (1-\pi)^m & , \quad n = 0 \\ \sum_{i=1}^{\min(n,m)}\binom{m}{i}\binom{n-1}{i-1}\left[\pi(1-\rho)\right]^i (1-\pi)^{m-i}\rho^{n-i} & , \quad n = 1, 2, \ldots. \end{cases} . \tag{14}$$

The mean and the variance of N are

$$\mathbb{E}[N] = \frac{m\pi}{1-\rho} \quad \text{and} \quad \mathbb{V}ar[N] = \frac{m\pi(1-\pi+\rho)}{(1-\rho)^2}, \tag{15}$$

respectively. The *p.g.f.* is given by

$$\mathbb{A}_N(z) = \left[1 - \frac{\pi(1-z)}{1-z\rho}\right]^m \quad \text{for} \quad 0 \le z \le 1, \tag{16}$$

while the long-term survival function of the CDB model is given by

$$S_{pop}(y) = \left[1 - \frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]^m. \tag{17}$$

Now, by letting $m \to \infty$ and $\pi \to 0$ in (17) such that $m\pi = \eta p > 0$, we obtain in the limit

$$\lim_{m\to\infty}\lim_{\pi\to 0} S_{pop}(y) = \lim_{m\to\infty}\left[1 - \frac{\eta p F(y)}{m(1-\rho(1-pF(y)))}\right]^m = \exp\left\{-\frac{\eta p F(y)}{1-\rho(1-pF(y))}\right\},$$

which is indeed the long-term survival function of the CDP model presented earlier in (12). If we set $m = p = 1$ and $\rho = 0$, the CDB model coincides with the mixture cure model of Boag (1949) and Berkson and Gage (1952).

3.3. Correlated destructive negative binomial (CDNB) model

For the choice of $a_n = \frac{\Gamma\left(\phi^{-1}+\sum_{i=1}^{\infty} n_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} n_i\right]!}$, $g(\theta) = (1-\theta)^{-\phi^{-1}}$ and parameter $\theta = \frac{\phi\eta}{1+\phi\eta}$, the number of initiated cells N has an inflated-parameter negative binomial distribution with parameters $\eta > 0$, $\rho \in [0,1)$, $\phi \ge -1$ and $\phi\eta > 0$, and its *p.m.f.* is of the form

$$\mathbb{P}_{NB}[N=n] = (1+\phi\eta)^{-\phi^{-1}} \sum_{n_1,n_2,\ldots} \frac{\Gamma\left(\phi^{-1}+\sum_{i=1}^{\infty} n_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} n_i\right]!}\left[\frac{\phi\eta(1-\rho)}{1+\phi\eta}\right]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{18}$$

where $n = 0, 1, \ldots$, and the summation is over all nonnegative integers $n_1, n_2, \ldots$, such that $\sum_{i=1}^{\infty} i n_i = n$, and $\Gamma(\cdot)$ denotes the gamma function. An alternate expression for the *p.m.f.* in (18) is given by

$$\mathbb{P}_{NB}[N=n] = \begin{cases} (1+\phi\eta)^{-\phi^{-1}} & , \quad n = 0 \\ (1+\phi\eta)^{-\phi^{-1}} \sum_{i=1}^{n} \binom{n-1}{i-1} \frac{\Gamma(\phi^{-1}+i)}{\Gamma(\phi^{-1})i!}\left[\frac{\phi\eta(1-\rho)}{1+\phi\eta}\right]^i \rho^{n-i} & , \quad n = 1,2,\ldots. \end{cases} \tag{19}$$

The mean and the variance of N are

$$\mathbb{E}[N] = \frac{\eta}{1-\rho} \quad \text{and} \quad \mathbb{V}ar[N] = \frac{\eta(1+\rho+\phi\eta)}{(1-\rho)^2}, \tag{20}$$

respectively. The *p.g.f.* is given by

$$\mathbb{A}_N(z) = \left[\frac{1-z\rho}{1+\phi\eta(1-z)-z\rho}\right]^{\phi^{-1}} \quad \text{for} \quad 0 \le z \le 1, \tag{21}$$

while the long-term survival function of the CDNB model is given by

$$S_{pop}(y) = \left[\frac{1-\rho(1-pF(y))}{1+\phi\eta pF(y)-\rho(1-pF(y))}\right]^{\phi^{-1}}. \tag{22}$$

When $\phi = 1$, we obtain the inflated-parameter geometric distribution with parameter $\theta = \frac{1}{1+\eta} \in (0,1)$ in (18) or (19), in which case $S_{pop}(\cdot)$ in (22) reduces to

$$S_{pop}(y) = \frac{1-\rho(1-pF(y))}{1+\eta pF(y)-\rho(1-pF(y))}, \tag{23}$$

giving rise to the correlated destructive geometric model, denoted simply by the CDG model. When $\phi \to 0$, we obtain the CDP model.

3.4. Correlated destructive logarithmic series (CDLS) model

For the choice of $a_n = \frac{(-1+n_1+n_2+\cdots)!}{n_1!n_2!\cdots}$, $g(\theta) = -\log(1-\theta)$ and $\theta = 1-\pi$, the number of initiated cells N has an inflated-parameter logarithmic series distribution with parameters $\pi \in (0,1)$ and $\rho \in [0,1)$, and its $p.m.f.$ is of the form

$$\mathbb{P}_{LS}[N=n] = (-\log(\pi))^{-1} \sum_{n_1,n_2,\ldots} \frac{(-1+n_1+n_2+\cdots)!}{n_1!n_2!\cdots} \big[(1-\pi)(1-\rho)\big]^{\sum_{i=1}^{\infty} n_i} \rho^{\sum_{i=2}^{\infty}(i-1)n_i}, \tag{24}$$

where $n = 0, 1, \ldots$, and the summation is over all nonnegative integers $n_1, n_2, \ldots$, such that $\sum_{i=1}^{\infty} i n_i = n$. An alternate expression for the $p.m.f.$ in (24) is given by

$$\mathbb{P}_{LS}[N=n] = (-\log(\pi))^{-1} \sum_{i=1}^{n} \binom{n-1}{i-1} \frac{\big[(1-\pi)(1-\rho)\big]^i \rho^{n-i}}{i}, \quad n = 1, 2, \ldots. \tag{25}$$

In its orginal form, this distribution excludes zero values due to which it cannot be used for modeling the number of initiated cells. For this reason, we consider here a modified inflated-parameter logarithmic series distribution, whose $p.m.f.$ can be expressed as

$$\mathbb{P}_{LS}[N=n] = (-\log(\pi))^{-1} \sum_{i=1}^{n+1} \binom{n}{i-1} \frac{\big[(1-\pi)(1-\rho)\big]^i \rho^{n+1-i}}{i}, \quad n = 0, 1, 2, \ldots. \tag{26}$$

The mean and the variance of the modified inflated-parameter logarithmic series random variable N are

$$\mathbb{E}[N] = 1 - \frac{1-\pi}{\pi(1-\rho)\log(\pi)} \quad \text{and} \quad \mathbb{V}ar[N] = -\frac{(1-\pi)\big[\log(\pi)(1+\pi\rho)+1-\pi\big]}{\pi^2(1-\rho)^2(\log(\pi))^2}, \tag{27}$$

respectively. The $p.g.f.$ is given by

$$\mathbb{A}_N(z) = \frac{(-\log(\pi))^{-1}}{z} \log\left\{\frac{1-\rho z}{1-z(1-\pi(1-\rho))}\right\}, \tag{28}$$

while the long-term survival function of the CDLS model is given by

$$S_{pop}(y) = \frac{(-\log(\pi))^{-1}}{(1-pF(y))} \log\left\{\frac{1-\rho(1-pF(y))}{1-(1-pF(y))(1-\pi(1-\rho))}\right\}. \tag{29}$$

In Table 3.1, we present the long-term survival function and the cured fraction corresponding to these specific models, as well as the improper density funtion $f_{pop}(y) = -S_{pop}(y)/dy$.

Table 3.1: Long-term survival function ($S_{pop}(y)$), density function ($f_{pop}(y)$), and cured fraction (p_0) for different special cases.

model	$S_{pop}(y)$	$f_{pop}(y)$	p_0
CDP	$\exp\left\{-\frac{\eta p F(y)}{1-\rho(1-pF(y))}\right\}$	$\left[\frac{\eta p f(y)\left[1-\rho(1-pF(y))\right]-\eta\rho p^2 f(y)F(y)}{\left[1-\rho(1-pF(y))\right]^2}\right] S_{pop}(y)$	$\exp\left\{-\frac{\eta p}{1-\rho(1-p)}\right\}$
CDB	$\left[1-\frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]^m$	$m\left[1-\frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]^{-1}\left[\frac{\pi p f(y)\left[1-\rho(1-pF(y))\right]-\pi p^2 F(y)\rho f(y)}{\left[1-\rho(1-pF(y))\right]^2}\right] S_{pop}(y)$	$\left[1-\frac{\pi p F(y)}{1-\rho(1-pF(y))}\right]^m$
CDNB	$\left[\frac{1-\rho(1-pF(y))}{1+\phi\eta p F(y)-\rho(1-pF(y))}\right]^{\phi^{-1}}$	$\phi^{-1}\left[\frac{1-\rho(1-pF(y))}{1+\phi\eta p F(y)-\rho(1-pF(y))}\right]^{-1}\left[\frac{\left[1-\rho(1-pF(y))\right]\left[\phi\eta p f(y)+\rho p f(y)\right]-\rho p f(y)\left[1+\phi\eta p F(y)-\rho(1-pF(y))\right]}{\left[1+\phi\eta p F(y)-\rho(1-pF(y))\right]^2}\right] S_{pop}(y)$	$\left[\frac{1-\rho(1-p)}{1+\phi\eta p-\rho(1-p)}\right]^{\phi^{-1}}$
CDLS	$\frac{(-\log(\pi))^{-1}}{(1-pF(y))}\log\left[\frac{1-\rho(1-pF(y))}{1-(1-pF(y))(1-\pi(1-\rho))}\right]$	$\frac{\left[1-(1-pF(y))(1-\pi(1-\rho))\right]}{\log(\pi)(1-\rho(1-pF(y)))(1-pF(y))}\left[\frac{\rho p f(y)}{1-(1-pF(y))(1-\pi(1-\rho))}-\frac{(1-\rho(1-pF(y)))p f(y)(1-\pi(1-\rho))}{\left[1-(1-pF(y))(1-\pi(1-\rho))\right]^2}\right]-\frac{p f(y)S_{pop}(y)}{1-pF(y)}$	$\frac{(-\log(\pi))^{-1}}{(1-p)}\log\left[\frac{1-\rho(1-p)}{1-(1-p)(1-\pi(1-\rho))}\right]$

4. Maximum likelihood estimation of model parameters

Let us consider the situation when the lifetime in (5) 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 $\delta_i = I(Y_i \leq C_i)$, where $\delta_i = 1$ if Y_i is a complete lifetime and $\delta_i = 0$ if it is right censored, for $i = 1, \ldots, n$. Let $\boldsymbol{\gamma}$ denote the parameter vector of the distribution of the unobserved time in (5). Note that the CDP, CDB and CDNB models in Table 3.1 are unidentifiable in the sense of Li et al. (2001). So, to circumvent this problem, we propose to relate the parameters p, η (or π) of the models in Table 3.1 to covariates $\boldsymbol{x_{1i}}$ and $\boldsymbol{x_{2i}}$, respectively, without common elements and $\boldsymbol{x_{2i}}$ without a column of ones. We adopt the link functions

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x_{1i}'\beta_1}, \quad \text{and} \quad \log(\eta_i) = \boldsymbol{x_{2i}'\beta_2} \quad \text{or} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_{2i}'\beta_2}, \quad i = 1, \ldots, n, \tag{30}$$

where $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ denote vectors with k_1 and k_2 coefficients.

A critical issue is the selection of covariates to be included in the link functions in (30). More precisely, given a link function and a set potential covariates, the problem is to find and fit the "best" model under a "selected" subset of covariates. Even though this problem is of importance, it will be not addressed here since in our illustrative example presented in the following section the covariates are just the ones already selected in the literature. For readers interested in this problem from the classical point of view, we suggest the books of Draper and Smith (1998) and Collet (1994).

From n pairs of times and censoring indicators $(t_1, \delta_1), \ldots, (t_n, \delta_n)$ as the available data, the likelihood function under non-informative censoring is given by

$$L(\boldsymbol{\vartheta}; \boldsymbol{t}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} \sum_{n_i=0}^{\infty} \sum_{d_i=0}^{n_i} \big[S(y_i; \boldsymbol{\gamma})\big]^{d_i - \delta_i} \big[d_i f(y_i; \boldsymbol{\gamma})\big]^{\delta_i} p(n_i, d_i; \theta_i, \rho_i, p) \tag{31}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, \boldsymbol{\gamma}', \rho, \phi, m))'$, $\boldsymbol{t} = (t_1, \ldots, t_n)'$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$. After some manipulations, the likelihood function in (31) can be expressed as

$$L(\boldsymbol{\vartheta}; \boldsymbol{t}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} \big[f_p(t_i; \boldsymbol{\vartheta})\big]^{\delta_i} \big[S_p(t_i; \boldsymbol{\vartheta})\big]^{1-\delta_i}, \tag{32}$$

where $f_{pop}(\cdot, \boldsymbol{\vartheta})$ and $S_{pop}(\cdot, \boldsymbol{\vartheta})$ for the models described in Section 3 are as presented in Table 3.1. We shall now assume a Weibull distribution for the unobserved lifetime in (5) with

$$F(z; \boldsymbol{\gamma}) = 1 - \exp(-z^{\gamma_1} e^{\gamma_2}) \quad \text{and} \quad f(z; \boldsymbol{\gamma}) = \gamma_1 z^{\gamma_1 - 1} \exp(\gamma_2 - z^{\gamma_1} e^{\gamma_2}) \tag{33}$$

for $z > 0$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)'$, with $\gamma_1 > 0$, and $\gamma_2 \in \Re$. Although other lifetime distributions can be considered here, our choice is due to the fact that the Weibull distribution is one of the most widely used lifetime distributions in survival analysis due to its flexibility. Depending on the value of its shape parameter γ_1, the Weibull distribution can be used to model a wide variety of failure rate behaviors. Its failure rate is monotone decreasing for $\gamma_1 < 1$, monotone increasing for $\gamma_1 > 1$, and constant for $\gamma_1 = 1$, leading to the exponential distribution. Furthermore, from a practical point of view, as shall be pointed out in the

next section, the failure rate for the melanoma data is indeed increasing and therefore can be properly modeled by a Weibull distribution.

From the likelihood function in (32), the maximum likelihood estimation of the parameter $\boldsymbol{\vartheta}$ can be carried out. Numerical maximization of the log-likelihood function $l(\boldsymbol{\vartheta}; \mathbf{t}, \boldsymbol{\delta}) = \log L(\boldsymbol{\vartheta}; \mathbf{t}, \boldsymbol{\delta})$ can be accomplished by using existing software (R Development Core Team, 2010). 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 ϑ and covariance matrix $\Sigma(\widehat{\boldsymbol{\vartheta}})$, which can be estimated by

$$\widehat{\Sigma}(\widehat{\boldsymbol{\vartheta}}) = \left[-\frac{\partial^2 l(\boldsymbol{\vartheta}; \boldsymbol{t}, \boldsymbol{\delta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'} \right]^{-1},$$

evaluated at $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$. The required second derivatives can be computed numerically once again by the use of R-software.

5. Application to a cutaneous melanoma data

The incidence of cutaneous malignant melanoma, a common cancer of the skin, is increasing dramatically in persons with light-colored skin in all parts of the world, being the second cause of potential life loss in years affecting younger adult individuals, only behind leukemia and causing a major public health problem (Barral, 2001).

In this section, we demonstrate an application of the models described in Section 3 to a cutaneous melanoma data. The dataset includes 205 patients observed after operation for removal of malignant melanoma in the period 1962-77. 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, with mean = 5.9 and standard deviation = 3.1 years), and refers to the time until the patient's death or the censoring time whichever came first. Patients who died from other causes as well as patients who were 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 here. Keeping in mind the identifiability issue mentioned earlier in Section 4, in the CDP, CDB and CDNB models the parameter p is linked only to tumor thickness, while the parameter η (or π) is linked to the ulceration status.

First, in order to identify the shape of a lifetime data hazard function, we used a graphical method based on the total time on test (TTT) plot (Aarset, 1985). In its empirical version, the TTT plot is given by $G(r/n) = [(\sum_{j=1}^{r} Y_{j:n}) + (n-r)Y_{r:n}]/(\sum_{j=1}^{r} Y_{j:n})$, where $r = 1, \ldots, n$ and $Y_{j:n}$ represent the order statistics of the sample. It has been shown that the hazard function increases (decreases) if the TTT plot is concave (convex). Although the TTT plot is only a sufficient condition, not a necessary one for indicating the shape of the hazard function, it is used here as a preliminary indication of its shape. Figure 5.1 (left panel) shows the TTT plot for the melanoma cancer data, which is concave indicating an increasing hazard function, which can be represented by a Weibull distribution. Kaplan-Meier curves, stratified by ulceration status (ulc), in Figure 5.1 (right panel), suggest cure levels of above 0.4. This indicates that models that ignore the possibility of cure will not be suitable for analyzing these data.

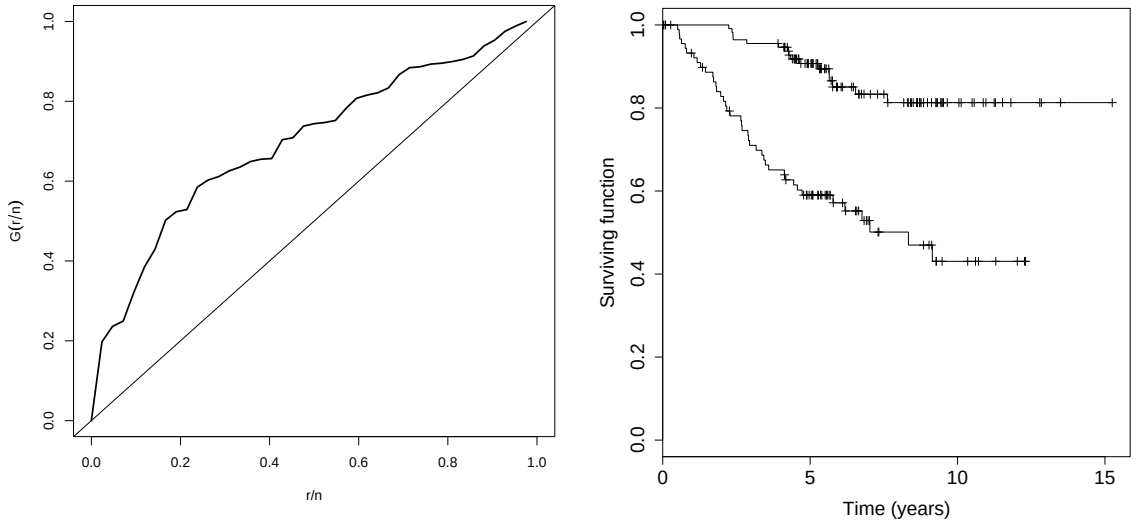

Figure 5.1: Left panel: TTTplot. Right panel: Kaplan-Meier curves stratified by ulceration status (upper: present, lower: absent).

Model comparison can be performed with the results presented in Table 5.1. Two particular cases of the CDNB model were also fitted to the data; namely, the negative binomial ($p = 1$, $\rho = 0$) and the geometric ($p = 1$, $\phi = 1$ and $\rho = 0$) models. In this way, the destruction mechanism is absent. For these models, the parameter η is linked to the two covariates. In order to compare the models, we used the $\max \log L(\cdot)$ values and the values of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which are defined, respectively, by $-2 \log L(\widehat{\boldsymbol{\vartheta}_g}) + 2q$ and $-2 \log L(\widehat{\boldsymbol{\vartheta}_g}) + q \log(n)$, where $\widehat{\boldsymbol{\vartheta}_g}$ is the MLE under the model g, q is the number of parameters estimated under the model g, and n is the sample size. The best model corresponds to a lower $\max \log L(\cdot)$, AIC and BIC values.

According to the above criteria, the CDG model stands out as the best one and so, we select the CDG model as our working model. Maximum likelihood estimates of the coefficients of the CDG model are presented in Table 5.2. The estimate of the correlation paramater (ρ) is 0.96, and as mentioned earlier in Section 2, this indicates a high association between cells. The estimate of the shape paramater (γ_1) provides an evidence against the exponential distribution ($\gamma_1 = 1$) for the unobserved lifetimes, corroborating what has been observed in the TTT plot in Figure 5.1.

Table 5.1: The values of $\max \log L(\cdot)$ and the AIC and BIC statistics for the seven fitted models, CDP, CDB, CDBN, CDG, CDLS, negative binomial and geometric models.

Criterion	CDP	CDB	CDNB	CDG	CDLS	Negative binomial	Geometric
$\max \log L(\cdot)$	-198.60	-198.61	-198.12	-198.52	-197.96	-201.52	-205.42
AIC	411.21	413.21	412.24	411.06	413.92	415.04	420.83
BIC	434.47	439.80	438.82	434.32	443.83	435.00	437.45

Table 5.2: Maximum likelihood estimates for the CDG model.

Parameter	**Estimate (est)**	**Standard error (se)**	**|est|/se**
γ_1	2.46	0.34	-
γ_2	-5.54	1.16	4.77
ρ	0.96	0.05	-
$\beta_{1,intercept}$	-4.84	0.95	5.10
$\beta_{1,thickness}$	0.95	0.27	3.55
$\beta_{2,ulc:absent}$	-0.48	0.41	1.17
$\beta_{2,ulc:present}$	0.53	0.30	1.76

Figure 5.2 displays the surviving function for patients with tumor thickness equal to 0.320, 1.94, and 8.32 mm, which correspond to the 5%, 50%, and 95% quantiles, respectively. The surviving probability is seen to decrease more rapidly for patients with thicker tumors. In Figure 5.2(a), the surviving function does not fall below 0.7.

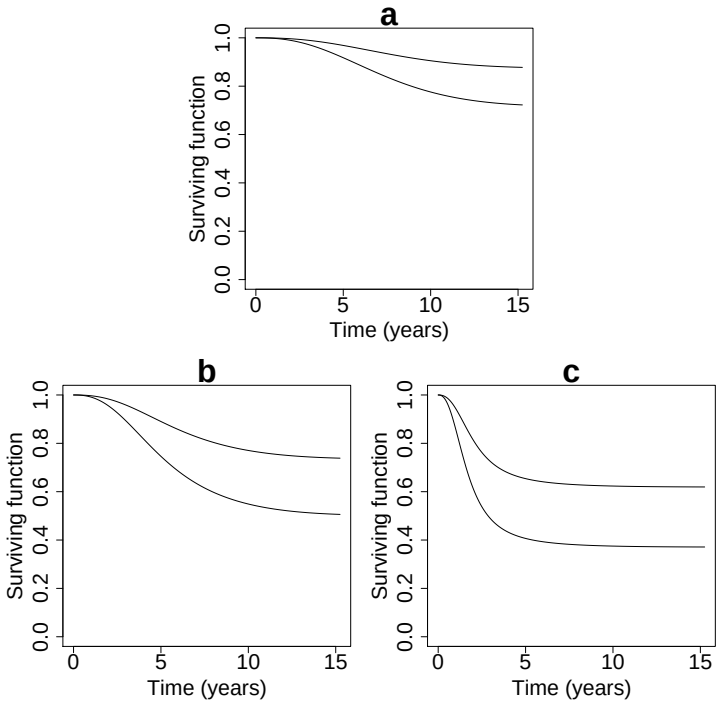

Figure 5.2: Surviving function under the IG model stratified by ulceration status (upper: absent, lower: present) for patients with tumor thickness equal to (a) 0.320, (b) 1.94, and (c) 8.32 mm, respectively.

The CDG model was fitted with the parameters p and η linked to tumor thickness and ulceration status, respectively. If we interchange these covariates, there is no improvement in the fit with respect to the criteria in Table 5.1, since in this case we obtain the values of ($\max \log L(\cdot)$, AIC, BIC) to be (-204.61, 423.23, 446.49).

Finally, we turn our attention to the role of the covariates on the cured fraction $p_0 = \frac{1-\rho(1-p)}{1+\eta p-\rho(1-p)}$. The estimates of the $\beta_{2,ulc}$ coefficients in Table 5.2 indicate that the mean number of initiated cells is greater

when ulceration is present, so that the cured fraction decreases. Since $\hat{\beta}_{2,thickness} > 0$ in Table 5.2, higher values of tumor thickness imply smaller cured fraction estimates. Figure 5.3 displays the combined effect of these covariates on the cured fraction. The lines run almost parallel and the cured fractions, after a steep decrease, for tumor thickness of over 5mm, are at 62.78% and 37.94% with ulceration status absent and present, respectively.

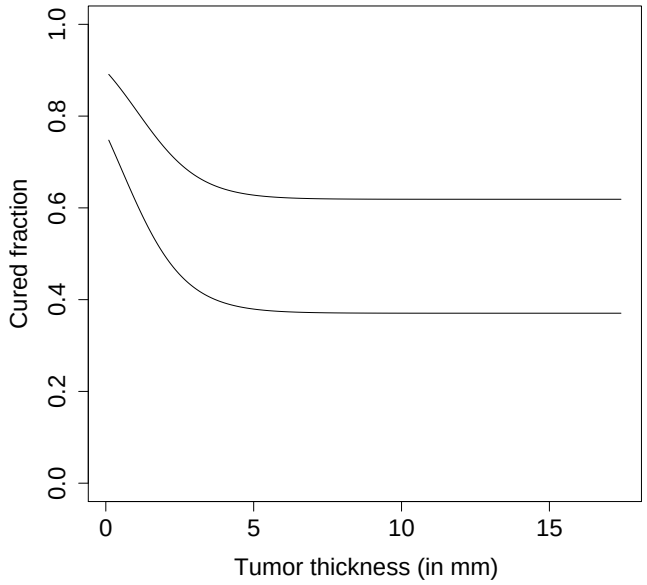

Figure 5.3: Cured fraction for the CDG model *versus* tumor thickness stratified by ulceration status (upper: absent, lower: present).

6. Concluding remarks

In this paper, we extend the models of Rodrigues et al. (2011) by proposing a model for survival data in the presence of latent dependent competing causes (tumour cells) and a cure fraction. We asssume a IGPS distribution for the number of initiated cells and a Weibull model for the lifetimes, and obtain the so-called CDGPS model. The CDGPS model incorporates into the analysis a biological dependence between the tumor cells. The advantage of this assumption is that we can measure the interdependence between the cells in an initiated tissue developing into a malignant tumor. The estimation procedure for the parameters of the proposed model is achieved through the maximum likelihood method. The practical relevance and applicability of the model are then demonstrated with a real data involving cutaneous melanoma. The proposed model, in addition to offering better interpretation to the underlying biological mechanism, offers better fit than the other commonly used cure rate models. The proposed CDGPS models will be quite helpful in the understanding of the biological process for a variety of infections, irradiation carcinogenesis and cancer chemoprevention experiments.

Acknowledgments: The authors express their sincere thanks to the Associate Editor and anonymous

referees for making some useful comments on an earlier version of this manuscript, which led to this improved version.

References

Aarset, M.V., 1985. The null distribution for a test of constant versus bathtub failure rate. Scandinavian Journal of Statistics 12, 55–68.

Banerjee, S., Carlin, B.P., 2004. Parametric spatial cure rate model for interval-cesnored time-to-relapse data. Biometrics 60, 268–275.

Barral, A.M., 2001. Immunological Studies in Malignant Melanoma: Importance of TNF and the Thioredoxin System. Doctorate Thesis, Linköping University, Linköping, Sweden.

Berkson, J., Gage, R.P., 1952. Survival curve for cancer patients following treatment. Journal of the American Statistical Association 42, 501–515.

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

Chen, M.H., Ibrahim, J.G., Sinha, D., 1999. A new bayesian model for survival data with a surviving fraction. Journal of the American Statistical Association 94, 909–919.

Chen, M.H., Ibrahim, J.G., Sinha, D., 2002. Bayesian inference for multivariate survival data with cure fraction. Journal of Multivariate Analysis 89, 101–126.

Collet, D., 1994. Modelling Survival Data in Medical Research. Chapman & Hall, London, UK.

Consul, P.C., 1990. New class of location-parameter discrete probability distributions and their characterizations. Communications in Statistics-Theory and Methods 19, 4653–4666.

Cooner, F., Banerjee, S., Carlin, B., Sinha, D., 2007. Flexible cure rate modelling under latent activation schemes. Journal of the American Statistical Association 102, 560–572.

Cox, D., Oakes, D., 1984. Analysis of Survival Data. Chapman & Hall, London, UK.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis. John Wiley & Sons, New York.

Farewell, V.T., 1982. The use of mixture models for the analysis of survival data with long term survivors. Biometrics 38, 1041–1046.

Farewell, V.T., 1986. Mixture models in survival analysis: Are they worth the risk? Canadian Journal of Statistics 14, 257–262.

Goldman, A.I., 1984. Survivorship analysis when cure is a possibility: A Monte Carlo study. Statistics in Medicine 3, 153–163.

Gordon, N.H., 1990. Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer patients. Statistics in Medicine 9, 397–407.

Gupta, R.C., 1974. Modified power series distributions and some of its applications. Sankhyã, Series B 35, 288–298.

Haynatzki, G.R., Weron, K., Haynatzka, V.R., 2000. A new statistical model of tumor latency time. Mathematical and Computer Modelling 32, 251–256.

Ibrahim, J.G., Chen, M.H., Sinha, D., 2001. Bayesian Survival Analysis. Springer-Verlag, New York.

Klebanov, L.B., Rachev, S.T., Yakovlev, A., 1993. A stochastic model of radiation carcinogenesis: Latent time distributions and their properties. Mathermatical Biosciences 113, 51–75.

Kolev, N. & Minkova, L..N.P., 2000. Inflated-parameter family of generalized power series distributions and their application in analysis of overdispersed insurance data. ARCH Research Clearing House 2, 295–320.

Li, C.S., Taylor, J., Sy, J., 2001. Identifiability of cure models. Statistics and Probability Letters 54, 389–395.

Minkova, L., 2002. A generalization of the classical discrete distributions. Communications in Statistics - Theory and Methods 31, 871–888.

Rodrigues, J., de Castro, M., Balakrishnan, N., Cancho, V.G., 2011. Destructive weighted Poisson cure rate models. Lifetime Data Analysis 17, 333–346.

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

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

Scheike, T., 2009. timereg package, with contributions from T. Martinussen and J. Silver. R package version 1.1-6.

Sy, J.P., Taylor, J.M.G., 2000. Estimation in a proportional hazards cure model. Biometrics 56, 227–336.

Yakovlev, A.Y., 1994. Parametric versus nonparametric methods for estimating cure rates based on censored survival-data. Statistics in Medicine 13, 983–985.

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

Yakovlev, A.Y., Tsodikov, A.D., Bass, L., 1993. A stochastic-model of hormesis. Mathematical Biosciences 116, 197–219.

Yang, G.L., Chen, C.W., 1991. A stochastic two-stage carcinogenesis model: a new approach to computing the probability of observing tumor in animal bioassays. Mathematical Biosciences 104, 247–258.

Yin, G., Ibrahim, J., 2005. A general class of bayesian survival models with zero and nonzero cure fractions. Biometrics 61, 403–412.