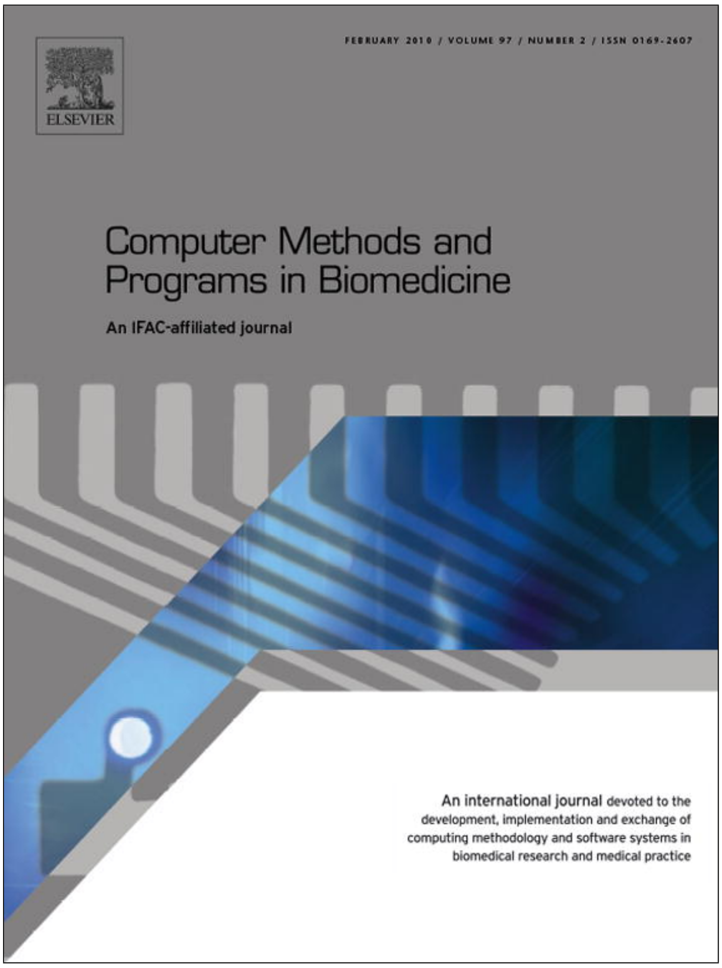
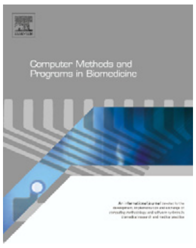

FEBRUARY 2010 / VOLUME 97 / NUMBER 2 / ISSN 0169-2607
ELSEVIER
Computer Methods and Programs in Biomedicine
An IFAC-affiliated journal
An international journal devoted to the development, implementation and exchange of computing methodology and software systems in biomedical research and medical practice

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



A hands-on approach for fitting long-term survival models under the GAMLSS framework

Mário de Castro[a,*], *Vicente G. Cancho*[a], *Josemar Rodrigues*[b]

[a] *Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, Caixa Postal 668, 13560-970, São Carlos-SP, Brazil*

[b] *Universidade Federal de São Carlos, Departamento de Estatística, Via Washington Luís, km 235, Caixa Postal 676, 13565-905, São Carlos-SP, Brazil*

ARTICLE INFO

Article history:
Received 19 October 2008
Received in revised form
5 August 2009
Accepted 11 August 2009

Keywords:
Survival analysis
Cure rate models
Long-term survival models
GAMLSS
Negative binomial distribution

ABSTRACT

In many data sets from clinical studies there are patients insusceptible to the occurrence of the event of interest. Survival models which ignore this fact are generally inadequate. The main goal of this paper is to describe an application of the generalized additive models for location, scale, and shape (GAMLSS) framework to the fitting of long-term survival models. In this work the number of competing causes of the event of interest follows the negative binomial distribution. In this way, some well known models found in the literature are characterized as particular cases of our proposal. The model is conveniently parameterized in terms of the cured fraction, which is then linked to covariates. We explore the use of the `gamlss` package in `R` as a powerful tool for inference in long-term survival models. The procedure is illustrated with a numerical example.

1. Introduction

Nowadays, due to advances in medical treatment and health care, in many data sets from clinical studies there are patients insusceptible to the occurrence of the event of interest (relapse of cancer or death, for instance), which can occur because of different competing causes. The proportion of such patients is termed the cured fraction. Models for survival data with a surviving fraction (also known as long-term survival models or cure rate models) occupy an outstanding place in reliability and survival analysis. The literature on the subject is by now vast and growing rapidly. The books by Maller and Zhou [1] and Ibrahim et al. [2], as well as the review article by Tsodikov et al. [3] and the article by Cooner et al. [4] could be mentioned as some key references.

In our paper the number of competing causes is modeled by the negative binomial distribution, leading to a formulation that encompasses some standard models found in the literature. Recently, studying cure rate models in the context of infectious diseases, Tournoud and Ecochard [5] proposed the negative binomial distribution for representing the number of competing causes. Distinct from these authors, we envision the long-term survival model with the negative binomial distribution as a general model encompassing the mixture cure model [6,7] and the promotion time cure model [8].

Another distinguishing facet of our work is the so-called Fisher's parameterization for the negative binomial distribution [9], meaning that the cured fraction plays a role of parameter in the model. Consequently, irrespective of the model, a unique expression relates the cured fraction to

[*] *Corresponding author*. Tel.: +55 16 33739567; fax: +55 16 33739751.
E-mail address: mcastro@icmc.usp.br (M. de Castro).

0169-2607/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2009.08.002

covariates. Previously this was never the case, for covariates are traditionally used to model the expectation of the number of competing causes (θ, say). The cured fraction is a function of θ and the relationship depends on the specific model. We argue that this uniqueness can be relevant in practice, since in many applications the cured fraction is the main quantity of interest and the practitioner would like to fit different models to the data with the same interpretation for the parameters of the models.

In order to put these concepts to work in practice we need a flexible modeling structure. Rigby and Stasinopoulos [10] presented a very broad class of models for a univariate response model, which they called GAMLSS. This acronym stands for generalized additive model for location, scale, and shape, represented by four parameters labeled as μ, σ, ν, and τ. The reader is referred to Rigby and Stasinopoulos [10] for a full account of the GAMLSS framework. Stasinopoulos and Rigby [11] described the implementation of GAMLSS in R [12]. We formulate our long-term survival model as an element in the GAMLSS class in order to take benefit of the gamlss package features.

In a spirit similar to our paper, Corbiére and Joly [13] proposed a SAS macro for parametric and semi-parametric mixture cure models. Our main contribution to the literature is twofold. First, the parametric model in the present paper is more general than the mixture cure model. Second, our computational platform makes use of software freely available.

The plan of the foregoing sections of the paper is as follows. Section 2 is dedicated to model formulation and parameter inference. Some details of the R functions are discussed in Section 3. Specifications and availability of the programs constitute the Sections 5 and 6. Results of an application to a real dataset are reported in Section 4. In Section 7 we end up with some general remarks.

2. Computational methods and theory

2.1. Model

The time to event for the j-th cause is denoted by Z_j, $j = 1, \ldots, M$, for $M \geq 1$, where $M \geq 0$ denotes the unobservable number of competing causes that can produce the event of interest. We assume that, conditional on M, the Z_j are i.i.d. with cumulative distribution function $F(t)$ and $S(t) = 1 - F(t)$. We assume also that M is independent of $Z_1, Z_2, \ldots$ The observable time to event is defined as $T = \min\{Z_1, \ldots, Z_M\}$, for $M \geq 1$, and $T = \infty$ if $M = 0$. Exponential, piecewise exponential [14], and Weibull distributions, for instance, can be used to represent Z_j. The i.i.d. assumption about $Z_1, Z_2, \ldots$ is surely a strong one, as remarked by Yakovlev and Tsodikov [8]. This option favors simplicity and analytical tractability at the expense of a more general formulation. Notwithstanding this limitation, such models have proven to be useful in many real-world applications. Under this setup, the survival function and the probability density function for the population are given, respectively, by

$$S_{\mathrm{pop}}(t) = P(M = 0) + \sum_{m=1}^{\infty} P(Z_1 > t, \ldots, Z_M > t | M = m) P(M = m)$$
$$= \sum_{m=0}^{\infty} S(t)^m \, P(M = m) \tag{1}$$

and

$$f_{\mathrm{pop}}(t) = -S'_{\mathrm{pop}}(t) = \sum_{m=0}^{\infty} m S(t)^{m-1} f(t) P(M = m),$$

where $f(t) = -S'(t)$ denotes the (proper) density function of the time to event Z in (1). Tsodikov et al. [3] and Rodrigues et al. [15], among others, proved that $S_{\mathrm{pop}}(t) = A_p(S(t))$, where $A_p(\cdot)$ is the probability generating function of the number of competing causes.

As in Tournoud and Ecochard [5], we suppose that the number of competing causes follows a negative binomial distribution with parameters θ and α [16,17], with probability function

$$P(M = m; \theta, \alpha) = \frac{\Gamma(\alpha^{-1} + m)}{m!\,\Gamma(\alpha^{-1})} \left(\frac{\alpha\,\theta}{1 + \alpha\,\theta} \right)^m (1 + \alpha\theta)^{-1/\alpha}, \tag{2}$$

$m = 0, 1, 2, \ldots$, for $\theta > 0$, $\alpha \geq -1$, and $1 + \alpha\theta > 0$. Negative values of α, $-1 \leq \alpha < 0$, lead to a range for m from 0 to the largest integer less than $-1/\alpha$ [9]. The expected value and the variance of M are

$$E(M) = \theta \quad \text{and} \quad \mathrm{var}(M) = \theta(1 + \alpha\theta). \tag{3}$$

The probability generating function is given by

$$A_p(s) = \sum_{m=0}^{\infty} P(M = m; \theta, \alpha)\, s^m = \{1 + \alpha\,\theta(1 - s)\}^{-1/\alpha},$$

$0 \leq s \leq 1$, so that the improper survival and density functions are given, respectively, by

$$S_{\mathrm{pop}}(t) = A_p(S(t)) = \begin{cases} \{1 + \alpha\theta F(t)\}^{-1/\alpha}, & \text{for } \alpha > -1/\theta,\ \alpha \neq 0, \\ \exp\{-\theta F(t)\}, & \text{for } \alpha = 0, \end{cases} \tag{4}$$

and

$$f_{\mathrm{pop}}(t) = -S'_{\mathrm{pop}}(t)$$
$$= \begin{cases} \{1 + \alpha\theta F(t)\}^{-1/\alpha - 1}\,\theta f(t), & \text{for } \alpha > -1/\theta,\ \alpha \neq 0, \\ \theta f(t) \exp\{-\theta F(t)\}, & \text{for } \alpha = 0. \end{cases} \tag{5}$$

The cured fraction is determined by $p_0 = \lim_{t \to \infty} S_{\mathrm{pop}}(t)$. From (2) or (4), $p_0 = (1 + \alpha\theta)^{-1/\alpha}$, for $\alpha > -1/\theta, \alpha \neq 0$, and $p_0 = \exp(-\theta)$, for $\alpha = 0$.

Besides the good fitting capabilities of the negative binomial model, its parameters have biological interpretations [5]. In (3) θ is the mean number of competing causes, whereas α accounts for the inter-individual variance of the number of causes. Additionally, the negative binomial distribution enabled de Castro et al. [18] to provide a probabilistic justification for the transformation introduced by Yin and Ibrahim [19].

As $\alpha \to 0$, we obtain the Poisson's distribution in (2) and $S_{\text{pop}}(t)$ in (4) gives rise to the promotion time cure model [8]. Regarding negative values of α in the $[-1, 0)$ interval, the negative binomial distribution still furnishes meaningful probabilities when $0 < -\alpha\theta < 1$ [9,16]. In particular, if $\alpha = -1$, $S_{\text{pop}}(t)$ in (4) becomes $S_{\text{pop}}(t) = 1 - \theta F(t)$, corresponding to the mixture cure model [6,7]. From (3) it follows that if $-1/\theta < \alpha < 0$, there is under-dispersion from the Poisson's model. On the other side, if $\alpha > 0$ the counts are over-dispersed.

2.2. Inference

From now on we suppose that the time to event is not completely observed and may be subject to right censoring. Let C_i denote the censoring time. We observe $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \leq C_i)$ is such that $\delta_i = 1$ if T_i is a time to event and $\delta_i = 0$ if it is right censored, $i = 1, \ldots, n$. Let $\boldsymbol{\gamma}$ denote the parameter vector of the distribution of the time to event Z in (1). From n pairs of times and censoring indicators $(y_1, \delta_1), \ldots, (y_n, \delta_n)$, the corresponding likelihood function under uninformative censoring is

$$L(\boldsymbol{\gamma}, \theta, \alpha; \mathbf{y}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} f(y_i, \delta_i; \boldsymbol{\gamma}, \theta, \alpha), \tag{6}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and

$$f(y_i, \delta_i; \boldsymbol{\gamma}, \theta, \alpha) = \sum_{m_i=0}^{\infty} S(y_i; \boldsymbol{\gamma})^{m_i - \delta_i} \{m_i f(y_i; \boldsymbol{\gamma})\}^{\delta_i}\, P(M = m_i; \theta, \alpha).$$

After some manipulations the likelihood function (6) can be written as

$$L(\boldsymbol{\gamma}, \theta, \alpha; \mathbf{y}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} f_{\text{pop}}(y_i; \boldsymbol{\gamma}, \theta, \alpha)^{\delta_i}\, S_{\text{pop}}(y_i; \boldsymbol{\gamma}, \theta, \alpha)^{1-\delta_i}. \tag{7}$$

Undoubtedly, in many instances the chiefest purpose of the analysis is the estimation of the cured fraction. With this concern in mind, we resort to the Fisher's parameterization of the negative binomial [9] in order to put the cured fraction p_0 in the expression of the likelihood function in (7). For $\alpha \geq -1$, we define $\theta = (p_0^{-\alpha} - 1)/\alpha$, if $\alpha \neq 0$, and $\theta = -\log(p_0)$, if $\alpha = 0$. Completing our model, we propose relate the cured fraction to covariates $\mathbf{x}_i$ by the logistic link, i.e.,

$$\log\left(\frac{p_{0i}}{1 - p_{0i}}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{or} \quad p_{0i} = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})}, \tag{8}$$

$i = 1, \ldots, n$, where $\boldsymbol{\beta}$ stands for the vector of regression coefficients. This parameterization is advantageous, for it enables one to think in the role of the covariates acting directly on the cured fraction within a familiar structure provided by the logistic link. Of course, probit and complement log–log link functions could also be mentioned and are available in the `gamlss` package.

From (3), $\text{var}(M_i) = E(M_i)\, p_{0i}^{-\alpha}$. Thus, extra variability in the number of competing causes due to omitted covariates is governed by the dispersion parameter α.

The improper functions in (4) and (5) are rewritten as

$$S_{\text{pop}}(y_i; \boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha) = \begin{cases} \{1 + (p_{0i}^{-\alpha} - 1)F(y_i; \boldsymbol{\gamma})\}^{-1/\alpha}, & \text{for } \alpha \geq -1,\ \alpha \neq 0, \\ p_{0i}^{F(y_i; \boldsymbol{\gamma})}, & \text{for } \alpha = 0, \end{cases} \tag{9}$$

and

$$f_{\text{pop}}(y_i; \boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha) = \begin{cases} \dfrac{p_{0i}^{-\alpha} - 1}{\alpha} \{1 + (p_{0i}^{-\alpha} - 1)F(y_i; \boldsymbol{\gamma})\}^{-1/\alpha - 1} f(y_i; \boldsymbol{\gamma}), & \text{for } \alpha \geq -1,\ \alpha \neq 0, \\ -\log(p_{0i})\, p_{0i}^{F(y_i; \boldsymbol{\gamma})} f(y_i; \boldsymbol{\gamma}), & \text{for } \alpha = 0. \end{cases} \tag{10}$$

Based on the negative binomial distribution, from (8), (9), and (10), the likelihood function in (7) is expressed by

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha; \mathbf{y}, \boldsymbol{\delta}) \propto \begin{cases} \displaystyle\prod_{i=1}^{n} \left\{\frac{p_{0i}^{-\alpha} - 1}{\alpha} f(y_i; \boldsymbol{\gamma})\right\}^{\delta_i} \left\{1 + (p_{0i}^{-\alpha} - 1)F(y_i; \boldsymbol{\gamma})\right\}^{-\delta_i - 1/\alpha}, & \text{for } \alpha \geq -1,\ \alpha \neq 0, \\ \displaystyle\prod_{i=1}^{n} \left\{-\log(p_{0i}) f(y_i; \boldsymbol{\gamma})\right\}^{\delta_i} p_{0i}^{F(y_i; \boldsymbol{\gamma})}, & \text{for } \alpha = 0. \end{cases} \tag{11}$$

Henceforth we assume a Weibull distribution for the time to event Z in (1). In our notation, $F(z; \boldsymbol{\gamma}) = 1 - \exp(-z^{\gamma_1} e^{\gamma_2})$ and $f(z; \boldsymbol{\gamma}) = \gamma_1 z^{\gamma_1 - 1} \exp(\gamma_2 - z^{\gamma_1} e^{\gamma_2})$, for $\gamma_1 > 0$ and $\gamma_2 \in \mathbb{R}$. In the absence of random effects, as in (8), parameter estimation is done through the maximum likelihood method applied to (11). The `gamlss` package requires the cumulative distribution function $F_{\text{pop}} = 1 - S_{\text{pop}}$ and the probability density function of the model, as in (9) and (10), and it does not matter whether they are proper or not.

2.2.1. Model identifiability

Identifiability of long-term survival models has attracted the attention of many researchers (e.g. [20,21,5]). Summing up the findings of these authors we conclude that our model is identifiable. Furthermore, following the steps in the proof of Theorem 6.2 in Tournoud and Ecochard [5], we conclude that if covariates are linked to the parameter $\tau = \alpha$ too, identifiability is preserved.

3. Program description

Our approach consists of an application of the `gamlss` function, which is fully documented in the `gamlss` package [11]. The cure rate models described in Section 2.2, specified by

Eqs. (9) and (10), are fitted using the `NBWEI4` and `POWEI4` functions. Many details about the `gamlss` function will be omitted from our presentation. The arguments of this function control the parameter estimation process. For instance, in the mixture cure model we have $\tau = \alpha = -1$. In a call to the `gamlss` function, this is passed as **tau.start** $= -1$ and **tau.fix** = TRUE. The promotion time cure model (`POWEI4` function) can also be fitted calling the `NBWEI4` function with `tau.start=0` and **tau.fix** = TRUE.

The long-term survival models we developed in Section 2 are implemented as a suite of `R` functions supplied in a text file (`models-cmpb.R`). This is the main file and contains the link function for the dispersion parameter α (remembering that $\alpha \geq -1$), the improper functions (9) and (10), the log-likelihood function corresponding to (11), *default* starting values for the parameters, and a function for construction of a confidence interval for the cured fraction. The structure of the `gamlss` function is familiar to readers used to the `R` (or S-Plus) syntax (the `glm` function, in particular).

4. Application

In this section we describe an example illustrating some tools we have developed with the `gamlss` package. This application was worked out by means of `R` commands stored in a text file (`example-cmpb.R`). The dataset includes 205 patients observed after operation for removal of malignant melanoma in the period 1962–1977. The patients were followed until 1977. These data are available in the `timereg` package [22]. The observed time (Y) 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. Patients dead from other causes, as well as patients still alive at the end of the study are censored observations (72%). The covariates are as follows: x_{i1}: ulceration (absent, $n = 115$; present, $n = 90$), x_{i2}: sex (female, $n = 126$; male, $n = 79$), and x_{i3}: tumor thickness (in mm, mean=2.92 and standard deviation=2.96), $i = 1, \ldots, 205$. For illustrative purposes, tumor thickness is categorized into two groups adopting 2 mm as cut point (median=1.94 mm). We are interested in the effect of the tumor thickness on the cured fraction. The parameterization of the Weibull distribution in Section 2.2 is named `WEI4` in our program.

As an example, the negative binomial WEI4 model is fitted issuing the command

```
gamlss(Surv(days, status) ~ 1, family = cens(NBWEI4),
 nu.formula = ~ thickness.le.2 + thickness.gt.2 - 1,
 data = datam, c.crit = 0.001, n.cyc = 1000)
```

Table 1 – Statistics from the adjusted models.

Model	Statistic		
	Global deviance	AIC	SBC
Negative binomial	414.4	424.4	441.0
Promotion time cure ($\alpha = 0$)	421.0	429.0	442.3
Mixture cure ($\alpha = -1$)	423.8	431.8	445.1

The `Surv` function applied to the observed time (`days`) and the censoring indicator (`status`) returns the censored times marked with a plus sign. In the `R` code the `NBWEI4` function encapsulates the elements of the model, which depends on the improper functions in (9) and (10). With the `cens` function in the `gamlss.cens` package we create a censored version of the distribution of the response variable, whose likelihood function is given by (11). The cured fraction ($\nu = p_0$) is linked to the categorized tumor thickness (`thickness.le.2` and `thickness.gt.2`) by the `nu.formula` specification using the logistic link in (8) by *default*. For the promotion time cure model the function is `POWEI4`.

The fitted models can be compared employing the log-likelihood function at its maximum (global deviance in the GAMLSS terminology, expressed by $-2 \max \log L$), the Akaike information criterion (AIC), and the Schwartz Bayesian criterion (SBC). Table 1 displays these statistics in increasing order of *AIC*. Negative binomial model yields the best fitting according to these criteria.

The worm plots (`wp` function) of the quantile residuals in Fig. 1 suggest that the negative binomial WEI4 is acceptable. Fig. 2 shows the Kaplan–Meier estimates of the survival function (solid lines) and estimates obtained from different parametric models. Kaplan–Meier curves level off above 0.8 and 0.4. Fig. 2(a) and (b) shows that the mixture cure and the promotion time cure models do not afford satisfactory fittings.

Taking into account the results in Table 1, Figs. 1 and 2, from this point on we select the negative binomial WEI4 model (m4, say) as our working model. Parameter estimates are obtained with the `summary(m4)` command, whose output is listed in Appendix A. All the coefficients are significant at a 1% level. The estimate of the shape parameter of the Weibull distribution (`WEI4`) is $e^{0.8716} = 2.4$ ($\hat{\mu} = \hat{\gamma}_1 =$`m4$mu.fv`). Comparing patients from tumor thickness categories "$\leq$ 2 mm" and "$>$ 2 mm", the odds ratio of the cured fraction ($\nu = p_0$) is estimated as `exp(m4$nu.coef[1] − m4$nu.coeff[2])` $= e^{1.609} = 5.0$. This estimate comes straightforwardly from the printout in Appendix A thanks to the parameterization in (11). For the negative

Table 2 – Estimates of the cured fraction and approximate confidence intervals from the negative binomial WEI4 model.

Tumor thickness category	Cured fraction	95% confidence interval		
		Asymptotic	Bootstrap	Deviance
$\leq$ 2 mm	0.803	(0.649, 0.900)	(0.592, 0.838)	(0.010, 0.890)
$>$ 2 mm	0.450	(0.327, 0.579)	(0.372, 0.584)	(0.010, 0.567)

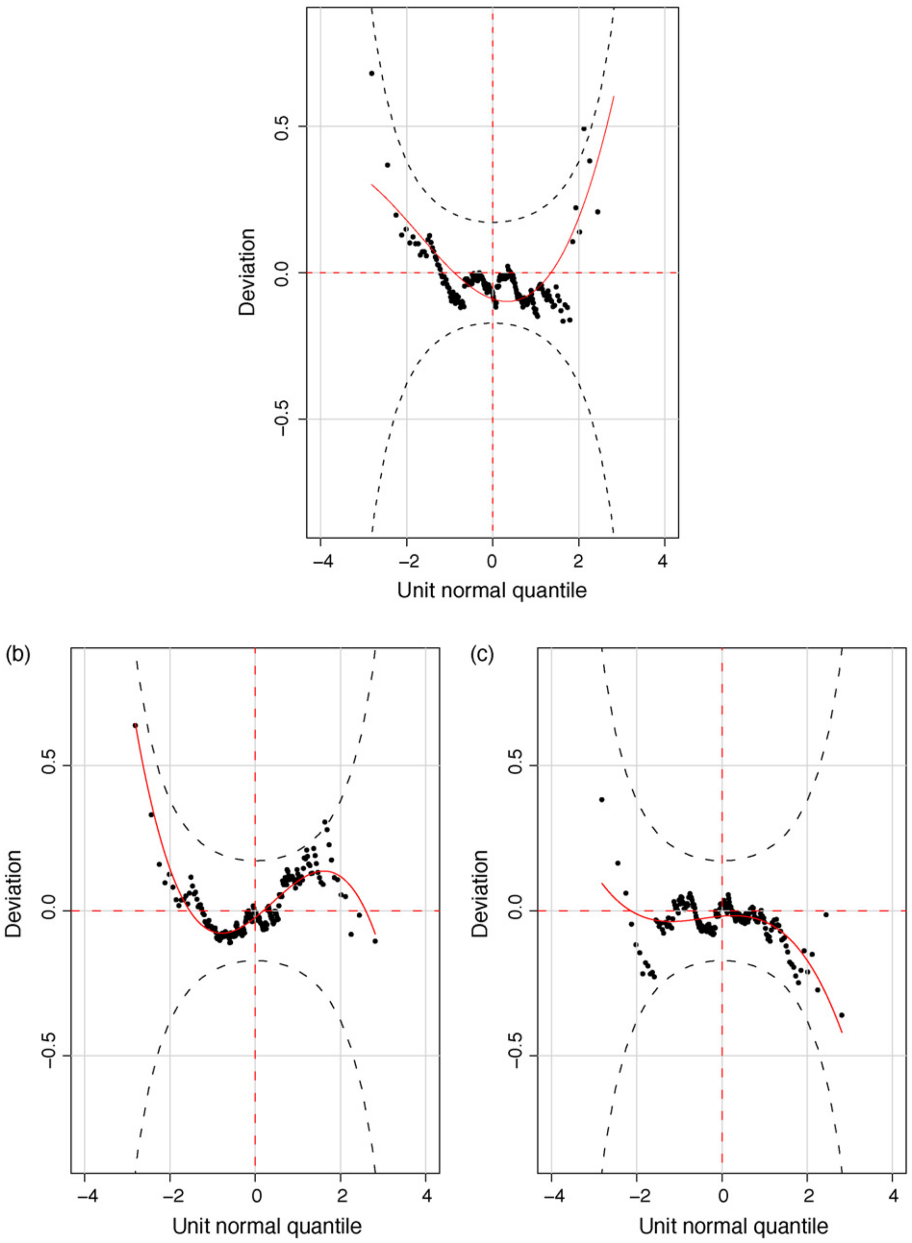


Fig. 1 – Worm plots of the quantile residuals: (a) mixture cure model, (b) promotion time cure model, and (c) negative binomial model.

binomial dispersion parameter ($\tau = \alpha$), the `own` link means that in our program the logarithm function was shifted to ensure that $\tau = \alpha \geq -1$. In this example, $\hat{\alpha}$ =`m4$tau.fv`= $e^{1.739} - 1 = 4.7$, since τ has a shifted log link function $\log(\tau + 1)$, which represents an evidence against the mixture cure and the promotion time cure models.

We end up our application dealing with the estimation of the cured fraction. Estimates of the cured fraction of patients stratified by tumor thickness category (and approximate 95% confidence intervals) are in Table 2. Asymptotic intervals are obtained from the output in Appendix A. In the computation of the intervals we used the `vcov` function (see the `example-cmpb.R` file). Bootstrap percentile intervals are based on 500 replicates with stratified case re-sampling adopting tumor thickness categories as strata. Bootstrap computations were performed using the `censboot` and `boot.ci` functions [23]. We notice that the differences between the intervals from these two methods are not so strong and the intervals for the patients of the two categories do not overlap. The thicker is the tumor, the lower is the cured fraction.

As a referee pointed out, a confidence interval for the cured fraction can also be obtained from a profile deviance curve. The confidence intervals for the parameter (β) in (8) por-

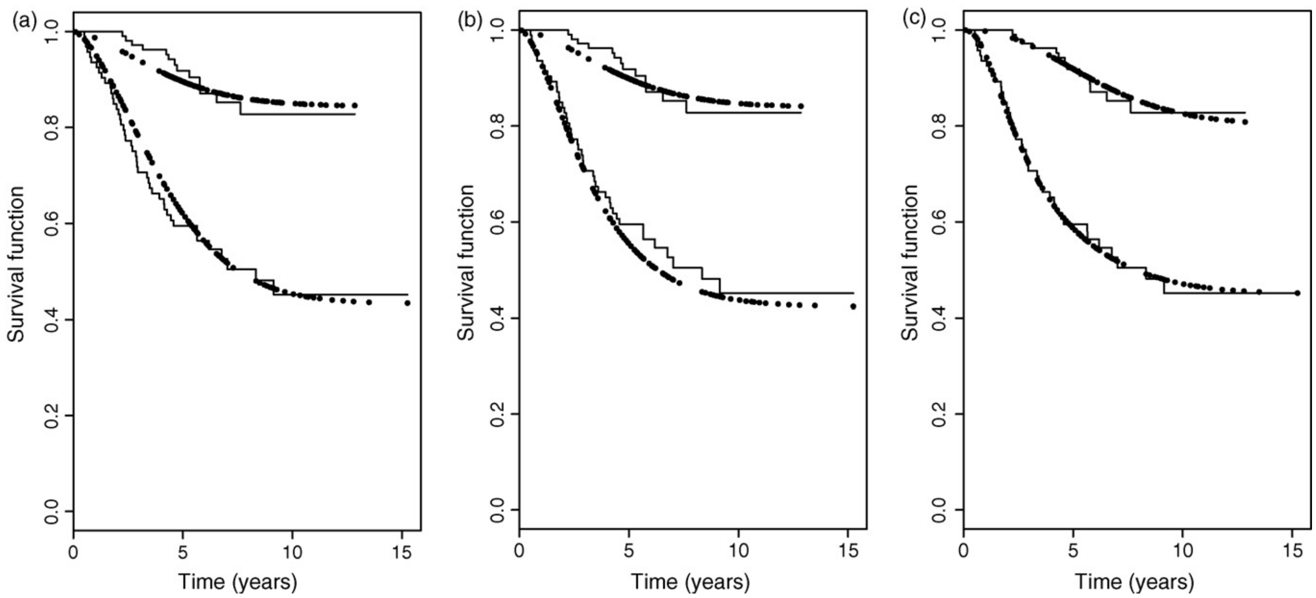


Fig. 2 – Kaplan–Meier curves stratified by tumor thickness category ("≤ 2 mm" and "> 2 mm", from top to bottom) and estimates of the survival function according to different WEI4 models: (a) mixture cure model, (b) promotion time cure model, and (c) negative binomial model.

trayed in Figs. 3 and 4 (patients with tumor thickness ≤ 2 and > 2 mm, respectively) were constructed with the `prof.term` function kindly supplied by an anonymous referee. Next, by applying the second expression in (8) the intervals for the cured fraction were calculated. In the last column of Table 2 the resulting intervals are very wide, suggesting confounding between the heaviness of the tail of the improper density function in (10) and the probability of a patient being cured. Hence, a more stable quantity of interest is the proportion of people who survived beyond a certain fixed time. For illustration, we choose 5 years. This proportion is estimated from (9) as $\hat{S}_{\text{pop}}(5) = S_{\text{pop}}(5; \hat{\gamma}, \hat{\boldsymbol{\beta}}, \hat{\alpha})$. Estimates of $S_{\text{pop}}(5)$ (and 95% bootstrap percentile confidence intervals) are 0.919 (0.825, 0.931) and 0.586 (0.562, 0.695) for tumor thickness ≤ 2 and > 2 mm, respectively. The density plots (`density` function) in Fig. 5 indicate that the effect of tumor thickness is more unambiguous on the proportion of patients who survived beyond 5 years than on the cured fraction.

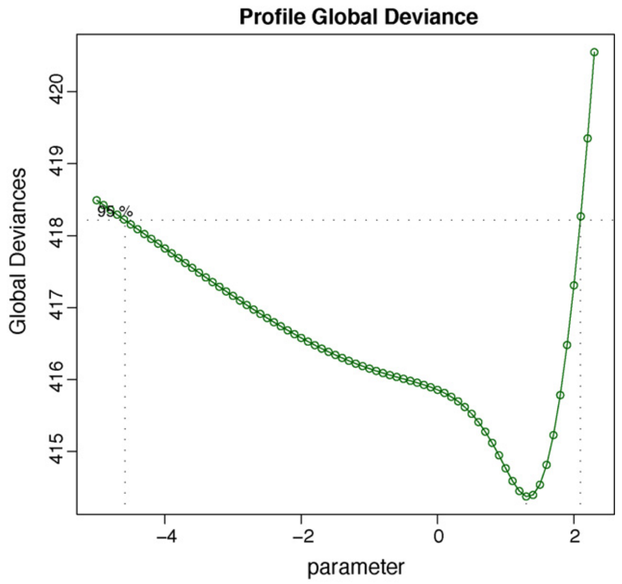


Fig. 3 – Profile deviance plot and 95% confidence interval for the parameter of the cured fraction – tumor thickness ≤ 2 mm.

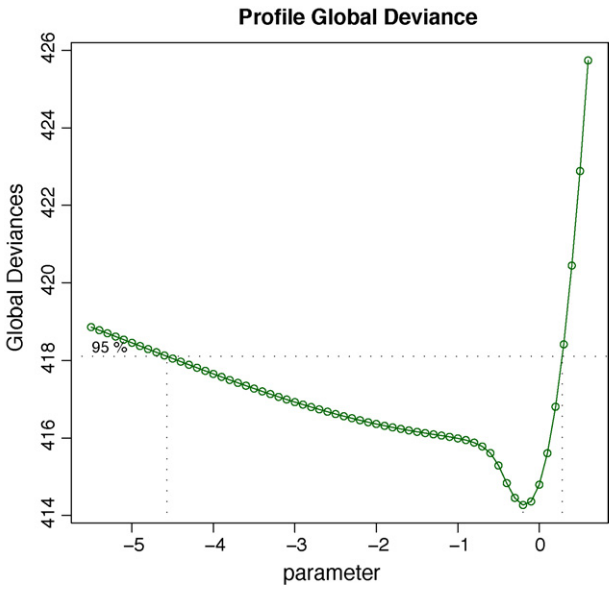


Fig. 4 – Profile deviance plot and 95% confidence interval for the parameter of the cured fraction – tumor thickness > 2 mm.

More complex models can be readily fitted. For instance, by running the commands

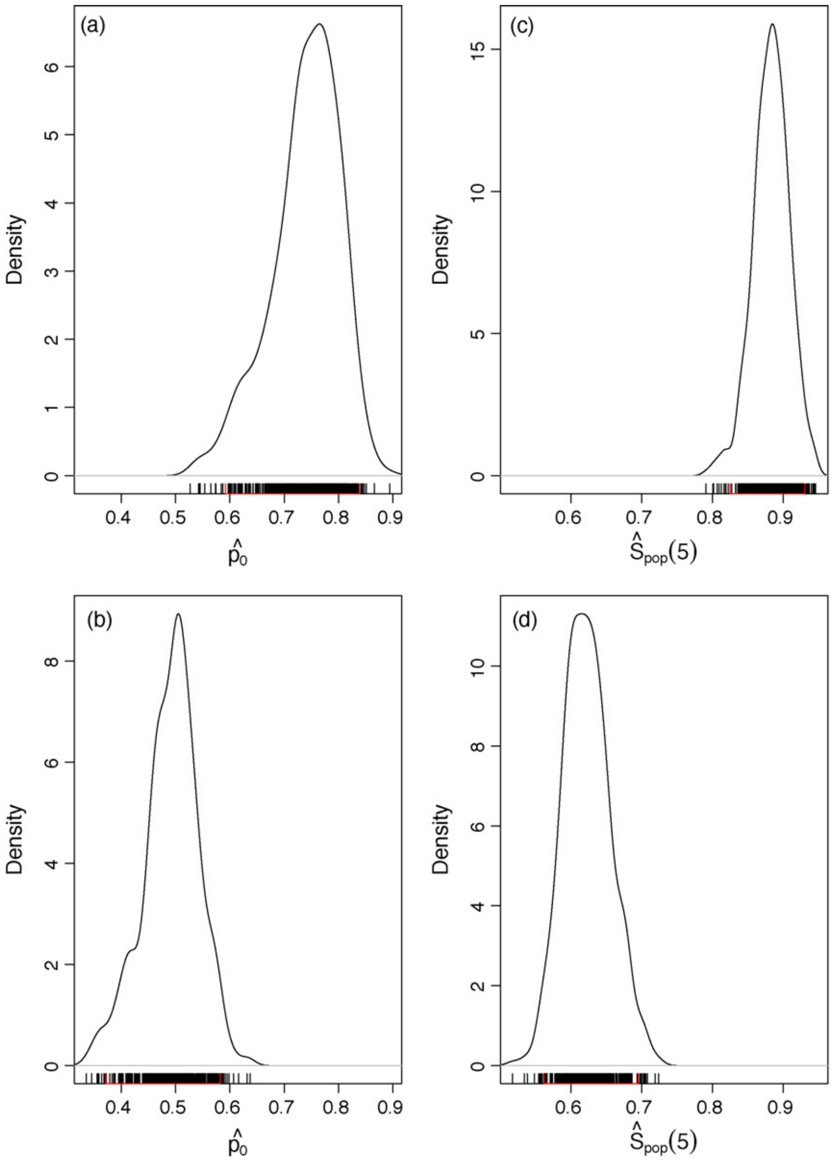


Fig. 5 – Density functions of bootstrap estimates of the cured fraction and the proportion of patients who survived more than 5 years together with 95% percentile confidence intervals (in red). (a) and (c) tumor thickness ≤ 2 mm, (b) and (d) tumor thickness > 2 mm. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

```
cat.thickness = cut(datam$thick, c(0, 2, max(datam$thick)),
  labels = c(" <= 2 mm", " > 2 mm"))
m4f = gamlss(Surv(days, status) ~ 1, family = cens(NBWEI4),
  nu.formula = ~ ulc + cat.thickness + sex,
  tau.formula = ~ ulc, data = datam, c.crit = 0.001, n.cyc = 1000,
  start.from = m4)
```

the negative binomial parameters ($\nu = p_0$ and $\tau = \alpha$) are linked to (ulceration, tumor thickness, and sex) and ulceration, respectively. Stepwise model selection can be done with the `stepGAIC` function.

5. Hardware and software specifications

The example in Section 4 was developed in a PC workstation under Microsoft Windows XP® and Ubuntu Linux operating systems.

6. Mode of availability

Computational codes of the long-term survival models and the example in Section 4 can be downloaded from http://www.icmc.usp.br/∼mcastro/download.html.

The R system, the gamlss, gamlss.cens, boot, and timereg packages, as well as information about the GAMLSS framework are freely available from Internet servers around the world. They are reachable at http://www.R-project.org and http://www.gamlss.com.

7. Conclusion

Under the negative binomial distribution for the number of competing causes, we present a unifying formulation of a long-term survival model. The parameterization in terms of the cured fraction distinguishes our paper from the usual proposals (e.g. [3,19,5]). Whichever the model, particularized by α in (9), covariates are related to the cure rate through the logistic link in (8). This issue can be attractive to practitioners.

Our example illustrated the possibility of trying out different models. Plots like the ones in Figs. 1 and 2 are a valuable tool for the model comparison task. Moreover, the GAMLSS framework is a powerful environment for the development of models not covered in this paper.

Conflict of interest

The authors have declared no conflict of interest.

Acknowledgements

The authors would like to thank two anonymous referees for their comments which contributed to improve the paper. In particular, one referee gave us detailed suggestions and supplied R code, including the prof.term function. The second and the third authors acknowledge partial support from CNPq, Brazil (grants number 473333/2008-2 and 300147/2008-2, respectively).

Appendix A.

Summary of the fitting of model m4 in Section 4:

```
Family:  c("NBWEI4rc", "right censored Neg. binom. WEI4 model")

Call:

gamlss(formula = Surv(days, status) ~ 1,

    nu.formula = ~ thickness.le.2 + thickness.gt.2 - 1,

    family = cens(NBWEI4), data = datam,

    c.crit = 0.001, n.cyc = 1000)

Fitting method: RS()

-------------------------------------------------------------------

Mu link function:  log

Mu Coefficients:

  Estimate   Std. Error     t value    Pr(>|t|)

 8.716e-01   2.116e-01   4.119e+00   5.576e-05

-------------------------------------------------------------------

Sigma link function:  identity

Sigma Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)

 -5.001514    1.275495   -3.921235    0.000121

----------------------------------------------------------------

Nu link function:  logit

Nu Coefficients:

                Estimate  Std. Error  t value   Pr(>|t|)

thickness.le.2    1.4066      0.4005   3.5118  0.0005501

thickness.gt.2   -0.2022      0.2645  -0.7646  0.4453973

----------------------------------------------------------------

Tau link function:  own

Tau Coefficients:

  Estimate  Std. Error     t value    Pr(>|t|)

 1.7391099   0.4396326   3.9558256   0.0001059

----------------------------------------------------------------

No. of observations in the fit:  205

Degrees of Freedom for the fit:  5

      Residual Deg. of Freedom:  200

                      at cycle:  202

Global Deviance:     414.4023

            AIC:     424.4023

            SBC:     441.0174
```

REFERENCES

[1] R.A. Maller, X. Zhou, Survival Analysis with Long-Term Survivors, Wiley, Chichester, UK, 1996.

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

[3] A.D. Tsodikov, J.G. Ibrahim, A.Y. Yakovlev, Estimating cure rates from survival data: an alternative to two-component mixture models, J. Am. Stat. Assoc. 98 (2003) 1063–1078.

[4] F. Cooner, S. Banerjee, B.P. Carlin, D. Sinha, Flexible cure rate modeling under latent activation schemes, J. Am. Stat. Assoc. 102 (2007) 560–572.

[5] M. Tournoud, R. Ecochard, Promotion time models with time-changing exposure and heterogeneity: application to infectious diseases, Biom. J. 50 (2008) 395–407.

[6] J.W. Boag, Maximum likelihood estimates of the proportion of patients cured by cancer therapy, J. R. Stat. Soc. B 11 (1949) 15–53.

[7] J. Berkson, R.P. Gage, Survival cure for cancer patients following treatment, J. Am. Stat. Assoc. 47 (1952) 501–515.

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

[9] G.J.S. Ross, D.A. Preece, The negative binomial distribution, Statistician 34 (1985) 323–336.

[10] R.A. Rigby, D.M. Stasinopoulos, Generalized additive models for location, scale and shape (with discussion), Appl. Stat. 54 (2005) 507–554.

[11] D.M. Stasinopoulos, R.A. Rigby, Generalized additive models for location scale and shape (GAMLSS) in R, J. Stat. Softw. 23 (2007) 1–46.

[12] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2009.

[13] F. Corbière, P. Joly, A SAS macro for parametric and semiparametric mixture cure models, Comput. Methods Programs Biomed. 85 (2007) 173–180.

[14] M.-H. Chen, J.G. Ibrahim, Maximum likelihood methods for cure rate models with missing covariates, Biometrics 57 (2001) 43–52.

[15] J. Rodrigues, V.G. Cancho, M. de Castro, F. Louzada-Neto, On the unification of the long-term survival models, Stat. Probabil. Lett. 79 (2009) 753–759.

[16] W.W. Piegorsch, Maximum likelihood estimation for the negative binomial dispersion parameter, Biometrics 46 (1990) 863–867.

[17] K. Saha, S. Paul, Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter, Biometrics 61 (2005) 179–185.

[18] M. de Castro, V.G. Cancho, J. Rodrigues, A note on a unified approach for cure rate models, Braz. J. Probabil. Stat., in press, available at http://www.imstat.org/bjps/future_papers.html.

[19] G. Yin, J.G. Ibrahim, Cure rate models: a unified approach, Can. J. Stat. 33 (2005) 559–570.

[20] C.S. Li, J.M.G. Taylor, J.P. Sy, Identifiability of cure models, Stat. Probabil. Lett. 54 (2001) 389–395.

[21] Y. Peng, J. Zhang, Identifiability of a mixture cure frailty model, Stat. Probabil. Lett. 78 (2008) 2604–2608.

[22] T. Scheike, Timereg package, with contributions from T. Martinussen and J. Silver, R package version 1.1–6, 2009.

[23] A. Canty, B. Ripley, Boot: bootstrap R (S-Plus) functions, R package version 1.2–33, 2008.