Lifetime Data Analysis
https://doi.org/10.1007/s10985-022-09575-3

A general class of promotion time cure rate models with a new biological interpretation

Yolanda M. Gómez[1] · Diego I. Gallardo[1,2] · Marcelo Bourguignon[3] · Eduardo Bertolli[4,5] · Vinicius F. Calsavara[6]

Received: 28 October 2021 / Accepted: 30 August 2022

Abstract

Over the last decades, the challenges in survival models have been changing considerably and full probabilistic modeling is crucial in many medical applications. Motivated from a new biological interpretation of cancer metastasis, we introduce a general method for obtaining more flexible cure rate models. The proposal model extended the promotion time cure rate model. Furthermore, it includes several well-known models as special cases and defines many new special models. We derive several properties of the hazard function for the proposed model and establish mathematical relationships with the promotion time cure rate model. We consider a frequentist approach to perform inferences, and the maximum likelihood method is employed to estimate the model parameters. Simulation studies are conducted to evaluate its performance with a discussion of the obtained results. A real dataset from population-based study of incident cases of melanoma diagnosed in the state of São Paulo, Brazil, is discussed in detail.

Keywords Compound Poisson distributions · Concurrent causes · Cure rate models · Promotion time cure model · Weibull distribution

✉ Marcelo Bourguignon
m.p.bourguignon@gmail.com

[1] Departamento de Medicina, Facultad de Medicina, Universidad de Atacama, Copiapó, Chile

[2] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile

[3] Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN 59078-970, Brazil

[4] Skin Cancer Department, A.C.Camargo Cancer Center, São Paulo, SP, Brazil

[5] Oncology Center, Beneficência Portuguesa, São Paulo, SP, Brazil

[6] Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center, Los Angeles, CA, USA

Published online: 16 September 2022

1 Introduction and motivation

In many clinical studies, there is a fraction of patients who are not susceptible to the occurrence of the event of interest. In this sense, long-term survival models play an important role and have been used for modeling time-to-event data, such as breast cancer and melanoma cancer. In a competing risk scenario, the promotion time cure rate (PTCR) model (Yakovlev and Tsodikov 1996) is one of the most important model in survival analysis, which has generated several theoretical (Yin and Ibrahim 2005; Chen and Du 2018) and practical (Tournoud and Ecochard 2007, 2008) researches focused on the PTCR model.

Motivated from a biological interpretation of cancer metastasis, Chen et al. (1999) studied the PTCR model with population survival function given by

$$S_{\text{pop}}(t;\lambda) = \exp\{-\lambda\,[1 - S(t)]\}, \quad t > 0,$$

where $\lambda > 0$, $\exp(-\lambda)$ is the long-term survivors or cured rate of the population, and $S(t)$ is a proper survival function of the susceptible patients. In this model, the number of concurrent causes (say N, a latent variable) is assumed to be a random variable following a Poisson distribution. However, the Poisson distribution is not always suitable, since its mean and variance are the same and this property may lead to bias in the estimate of the cure probability (Tucker et al. 1990). In the literature, other authors have considered different probability distributions for the latent variable. For instance, Rodrigues et al. (2009) studied the Conway–Maxwell–Poisson cure rate survival models, while Rodrigues et al. (2011) introduced the weighted Poisson cure rate models, and Rodrigues et al. (2016) proposed the relaxed Poisson cure rate models, among others. Such distributions assumed an additional parameter in order to incorporate underdispersion and overdispersion in the distribution of N.

In this paper, we extend the long-term survival model proposed by Chen et al. (1999) allowing a flexible family of cure rate models (many special cases), which is characterized by a new natural biological motivation of metastatic cancer. The new approach gives an interesting and realistic interpretation of the biological mechanism for the occurrence of the event of interest, which assumes the existence of a random number of clusters, each one of them with at least one cell. We proposed a cure rate model to accommodate the characteristics of unobservable stages of carcinogenesis from lifetime data in the presence of latent concurrent causes. In particular, we assume that the number of concurrent causes follows a compound Poisson distribution.

The main contributions and advantages of the proposed cure rate model are as follows: (1) Natural motivation: it is derived from a natural biological motivation of metastatic cancer (see Fig. 2); (2) Flexibility: the compound Poisson distribution is able to capture overdispersed and equidispersed activated cells; (3) Mathematical simplicity: its density and distribution functions have a simple form and do not involve complicated normalizing constants and/or special functions (see Eq. 3); (4) Special cases: it includes several well-known models as special cases and defines many new special models, such as the Hermite, Neyman type A, Thomas, Pólya–Aeppli, discrete stable, Poisson–inverse Gaussian, Poisson–Pascal and Poisson–Tweedie models, among others. See Table 3 in Wimme and Altman (1996). Furthermore, all special cases have the same cure rate; (5) Easy interpretation: it is indexed using by mean of the time-to-event for the concurrent causes and long-term survivors; (6) Double regression model: we allow a regression structure on the mean of the time-to-event for the concurrent causes and the long-term survivors' parameters (see Eq. (7)). Thus, we obtain a straightforward interpretation of the regression coefficients in terms of the expectation of the time-to-event for the cells and the long-term survivors; (7) Model estimation: estimation and inference are based on the likelihood paradigm (parametric approach), which can be easily computed using the R programming language (R Core Team 2020) through of the gamlss environment (Rigby and Stasinopoulos 2005). Therefore, our model can be easily used by researchers in several areas (see Sect. 3); (8) Applications: It showed a good performance in the extensive simulation studies and the applicability using a real dataset (see Sect. 4).

1.1 Motivation: melanoma cancer diagnosed in the state of São Paulo

In public policies and healthcare providers the clinical outcomes are fundamental. In oncology generally the survival rates, as cancer-specific survival, and/or disease-free survival rates is the researchers' main interest. The estimates for such survival rates can be obtained based on the cancer type and patient features, such as the age at diagnosis, sex, education level, clinical stage of the disease, type of treatment, and other available information in medical records. The melanoma-specific survival rates may vary from 24% to 88% (Gershenwald et al. 2017) after ten years. In Brazil, approximately 6000 new cases of melanoma were expected according to the Brazilian National Institute of Cancer (Coordenação de Prevenção e Vigilância 2017); and 7000 according to the International Agency for Research on Cancer (IARC) (Ervik et al. 2016); whereas, approximately 2000 deaths per year are attributable to melanoma in Brazil (Gershenwald et al. 2017; Ervik et al. 2016).

Considering tumor biology of melanomas, it has been demonstrated that some tumors may arise from benign lesions (melanocytic naevi). The predominant pathogenic mechanism that drives the progression of these naevi to melanoma is ultraviolet (UV) radiation mutagenesis. Other genetic mutations such as $BRAFV^{600E}$ are also found, especially in cases where there is no association to sun-damage (Shain and Bastian 2016). The concept of different cell clusters may be applied in this context. A greater probability of a long-term survivors is expected when the melanoma cancer is detected in early stages due to treatment as radical treatment, including surgery. In the routine clinical patients diagnosed in the clinical stages I or II are treated with surgery, and most of them will be alive after ten years of follow-up, while patients with clinical stages advanced other therapies are conducted, and its prognosis is worse due to its potential for metastatic dissemination. In skin cancer, in special melanoma, the patient's death can be attributed to different latent concurrent causes such as the presence of an unknown number of cancer cells in different clusters.

Even for stage IV melanomas, response rates after systemic treatments are still variable. It is known that metastatic cells may be very heterogeneous among themselves, which would lead to mixed (Berrino 2021) or partial (Puglisi 2021) responses patterns which can be also considered in the different cell cluster concept.

In our study, patients diagnosed with melanoma cancer were enrolled between 2000 and 2014 with follow-up conducted until 2018. All patients were followed after diagnosed and the death due to cancer was defined as the event of interest. Death due other causes or lost of follow-up were considered as been right-censored observations. It is part of a study of skin cancer in 6749 patients diagnosed melanoma in the state of São Paulo, Brazil. This dataset was initially studied by Calsavara et al. (2020), which they considered only an observed covariate (surgery) in the modeling. Recently, Molina et al. (2021), Rodrigues et al. (2021), Leão et al. (2021) and Gómez et al. (2021) considered other covariates available in the medical records, such as sex, clinical stage, and type of treatment (radiotherapy and chemotherapy). Our aim was to evaluate the effect of all explanatory variables measured at baseline, such as sex, age at diagnosis, clinical stage, surgery, radiotherapy, and chemotherapy in both components (failure rate and long-term survivors).

The dataset is from a retrospective survey of 6749 records of patients diagnosed with melanoma, of whom 3415 (51%) were female patients, the mean age was 58.04 (standard deviation = 16.36), and 4552 (67%) were in clinical stage I or II. Regarding treatment, 5978 (89%) patients underwent surgery, 587 (9%) patients received radiotherapy, and 1104 (16%) patients received chemotherapy. A total of 1912 (28%) events occurred during the follow-up period. The maximum observation time was approximately 18.54 years, while the median follow-up time was 5.24 years.

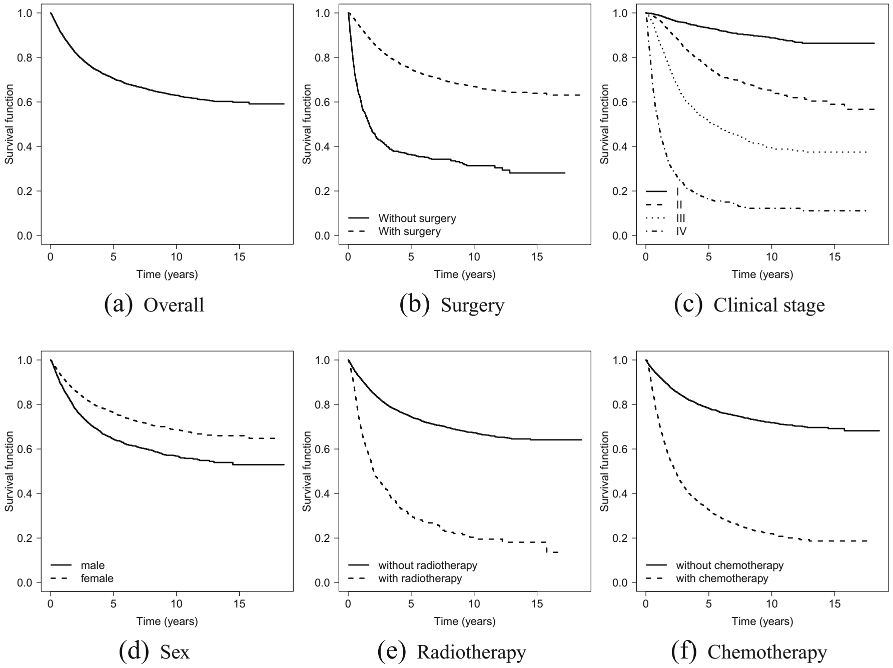
The estimated overall survival rates and stratified by surgery, clinical stage, sex, radiotherapy and chemotherapy were obtained by the Kaplan–Meier (KM) estimator and are shown in Fig. 1. The estimated curves suggest evidence of long-term survivors, regardless of the baseline characteristics. Among all of the variables considered in the study, those with clinical stage I melanoma had a better prognosis, as expected. The estimated overall 1-, 2-, 5-, and 10-year specific survival rates are 0.896, 0.820, 0.706 and 0.629, respectively.

According to the estimated survival curves (Fig. 1), each observed covariate is associate with the time-to-event and the long-term survivors. Therefore, in order to take into account the observed heterogeneity among the patients, we propose a new cure rate model in the next section.

The rest of the paper is organized as follows. In Sect. 2, we introduce and motivate the new cure rate models and discuss some of its properties. Furthermore, some specific models obtained as special cases from the general model are detailed. In Sect. 3, we describe the maximum likelihood estimation procedure. A Monte Carlo simulation is presented in Sect. 4 in order to evaluate the finite-sample behavior of the maximum likelihood estimators. A real data set from skin cancer also is discussed in Sect. 4. Finally, some conclusions are given in Sect. 5.

2 The new cure rate model

In this section, we introduce the new cure rate model, its main properties and some especial cases.



(a) Overall (b) Surgery (c) Clinical stage

(d) Sex (e) Radiotherapy (f) Chemotherapy

Fig. 1 Estimated survival curve obtained via Kaplan–Meier estimator for melanoma dataset for the overall, surgery, clinical stage, sex, radiotherapy, and chemotherapy

2.1 Formulation

Let N the number of clusters of cells for an individual left active after the initial treatment, and conditional on $N = n \geq 1$, Υ_j, $j = 1, \ldots, N$, be independent and identically distributed variables with the range being contained in $\mathbb{N} = \{1, 2, \ldots\}$, independent of N, indicating the number of cells in jth cluster, with probability mass function $\Pr(\Upsilon_j = i; \tau) = \tau_i$. Consider D be the total number of malignant cells (not eliminated by the treatment) defined by

$$D = \begin{cases} \sum_{j=1}^{N} \Upsilon_j, & \text{if } N \geq 1, \\ 0, & \text{if } N = 0. \end{cases}$$

We assume that the number of clusters of cells (N) follows a Poisson distribution with mean λ. In this case, the random variable D is said to have a compound Poisson (CP) distribution, where $\Pr(D = 0) = \mathrm{e}^{-\lambda}$. The mean and variance of D are, respectively

$$\mathrm{E}(D) \equiv \mu_D = \mathrm{E}(N)\mathrm{E}(\Upsilon) \quad \text{and} \quad \mathrm{Var}(D) \equiv \sigma_D^2 = \mathrm{E}(N)\mathrm{E}(\Upsilon^2).$$

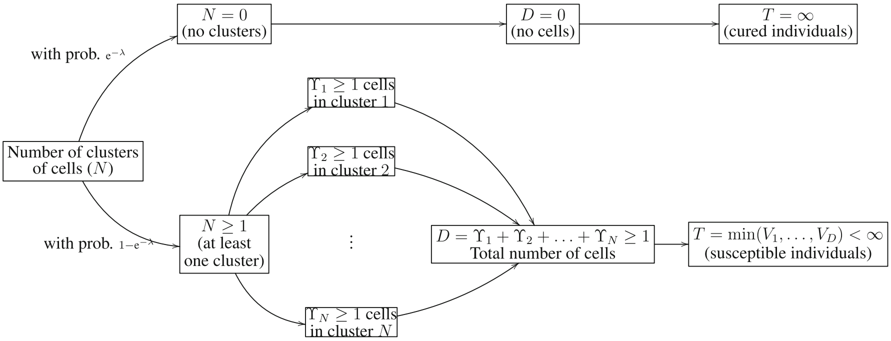

$N = 0$ (no clusters)
$D = 0$ (no cells)
$T = \infty$ (cured individuals)
with prob. $e^{-\lambda}$
$\Upsilon_1 \geq 1$ cells in cluster 1
Number of clusters of cells (N)
$\Upsilon_2 \geq 1$ cells in cluster 2
with prob. $1-e^{-\lambda}$
$N \geq 1$ (at least one cluster)
$\vdots$
$D = \Upsilon_1 + \Upsilon_2 + \ldots + \Upsilon_N \geq 1$ Total number of cells
$T = \min(V_1, \ldots, V_D) < \infty$ (susceptible individuals)
$\Upsilon_N \geq 1$ cells in cluster N

Fig. 2 Representation of the proposed model in a diagrammatic form

The dispersion index of the compound Poisson distribution is given by

$$\frac{\sigma_D^2}{\mu_D} = 1 + \frac{\mathrm{E}(\Upsilon(\Upsilon - 1))}{\mathrm{E}(\Upsilon)},$$

which means that the compound Poisson distribution is an over-dispersed model. The distribution of D is equi-dispersed if and only if $\mathrm{E}(\Upsilon^2) = \mathrm{E}(\Upsilon)$.

Denote the probability generating function (pgf) of the Υ_i (compounding distribution) by $\varphi_\Upsilon(s) := \mathrm{E}(s^\Upsilon) = \sum_{j=1}^{\nu} s^j \tau_j$, where $\sum_{j=1}^{\nu} \tau_j = 1$ and ν denotes the upper limit of the range (it allows the case $\nu = \infty$). In the case of a finite range with upper limit $\nu < \infty$ and $\tau_\nu > 0$, the expression of the pgf is reduced to $\varphi_\Upsilon(s) = \tau_1 s + \cdots + \tau_\nu s^\nu$. Then, the pgf of D, denoted by $\varphi_D(s)$, is given by

$$\begin{aligned}\varphi_D(s) = \varphi_N(\varphi_\Upsilon(s)) = e^{\lambda(\varphi_\Upsilon(s)-1)} &= e^{\lambda\left(\sum\limits_{j=1}^{\nu} s^j \tau_j - 1\right)} \\ = e^{\lambda\left(\sum\limits_{j=1}^{\nu} (s^j \tau_j - \tau_j)\right)} &= e^{\lambda\left(\sum\limits_{j=1}^{\nu} \tau_j (s^j - 1)\right)}. \end{aligned} \qquad (1)$$

A random variable D with pgf (1) is denoted by $\mathrm{CP}_\nu(\lambda, \varphi)$ (possibly $\nu = \infty$).

We assume that the $V_1, V_2, \ldots,$ are independent and identically distributed random variables representing the promotion times of the concurrent causes, independent of D, with survival function (SF) $S(\cdot; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. $S(\cdot; \boldsymbol{\theta})$ is proper in the sense that $\lim_{t\to\infty} S(t; \boldsymbol{\theta}) = 0$. In the concurrent causes scenario, the number of clusters of cells (N), the number of cells in jth cluster (Υ_j), the total number of cells (D), and the times V_k are all unobservable variables (latent variables). Thus, the observable time-to-event (which is the event of interest) in a given individual is defined by the random variable $T = \min\{V_1, \ldots, V_D\}$ for $D \geq 1$, and $T = \infty$ if $D = 0$. Figure 2 illustrates this interpretation.

Remark 2.1 When $\Upsilon_1 = \Upsilon_2 = \cdots = \Upsilon_N = 1$ the proposed model is reduced to the PTCR model (Yakovlev and Tsodikov 1996; Chen et al. 1999).

In this work, we assumed the minimum among $V_1, V_2, \ldots,$ because from a biological point of view, it seems appropriated that one carcinogenic cell can trigger the metastatic process. However, following the idea discussed in Kim et al. (2011) the model should be considered in a more general framework based on the threshold cure rate (TCR). In this scheme, it is assumed the existence of a random variable $R \leq 1$, independent from $V_1, V_2, \ldots,$. In simple words, R denotes the threshold to an individual to be considered as cured. The particular cases $R = 1$, $R = N$ and R with discrete uniform distribution on $1, \ldots, N$ were discussed in Cooner et al. (2007) and are known in the literature as the first activation, last activation and random activation schemes, respectively. However, Kim et al. (2011) also discuss other general cases where the distribution of R not depend on N. For instance, R fixed at an arbitrary value (say R_0), with shifted geometric or the shifted Poisson distributions.

Under this setup, the population SF, $S_{\text{pop}}(\cdot)$, can be computed as

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = \varphi_N(\varphi_\Upsilon(S(t;\boldsymbol{\theta}))) = \exp\left\{-\lambda\left[1-\sum_{j=1}^{\nu}\tau_j\left(S(t;\boldsymbol{\theta})^j\right)\right]\right\}, \quad t>0, \tag{2}$$

which depends on a parameter vector $\boldsymbol{\xi} = (\boldsymbol{\theta}, \lambda, \tau)^\top$ and $S(\cdot;\boldsymbol{\theta})$ denotes the SF of the promotion times of the concurrent causes, which is a proper function in the sense that $S(\infty;\boldsymbol{\theta}) = 0$. We also assumed that $\mu \subset \boldsymbol{\theta}$ represents the mean of this distribution (mean of the time-to-event of the carcinogenic cells), which must exist. The population density and population hazard rate (HR) functions are given by

$$f_{\text{pop}}(t;\boldsymbol{\xi}) = \lambda\, f(t;\boldsymbol{\theta})\sum_{j=1}^{\nu} j\,\tau_j\left(S(t;\boldsymbol{\theta})^{j-1}\right)\exp\left\{-\lambda\left[1-\sum_{j=1}^{\nu}\tau_j\left(S(t;\boldsymbol{\theta})^j\right)\right]\right\}, \quad t>0, \tag{3}$$

and

$$h_{\text{pop}}(t;\boldsymbol{\xi}) = \lambda\, f(t;\boldsymbol{\theta})\sum_{j=1}^{\nu} j\,\tau_j\, S(t;\boldsymbol{\theta})^{j-1} = \lambda\, h(t;\boldsymbol{\theta})\sum_{j=1}^{\nu} j\,\tau_j\, S(t;\boldsymbol{\theta})^{j}, \quad t>0, \tag{4}$$

where $f(t;\boldsymbol{\theta}) = -\mathrm{d}S(t;\boldsymbol{\theta})/\mathrm{d}t$ denotes the (proper) PDF of the time-to-event T in (2) and $h(t;\boldsymbol{\theta}) = f(t;\boldsymbol{\theta})/S(t;\boldsymbol{\theta})$ is the proper HR function of the time-to-event T.

Proposition 2.1 *It follows from* (4) *that*

(i) $h_{\text{pop}}(t;\boldsymbol{\xi}) = h_{\text{prom}}(t;\boldsymbol{\xi})\,\mathrm{E}(\Upsilon\, S(t;\boldsymbol{\theta})^{\Upsilon-1})$,
(ii) $h_{\text{pop}}(t;\boldsymbol{\xi}) \geq h_{\text{prom}}(t;\boldsymbol{\xi})$,

where $h_{\text{prom}}(t;\boldsymbol{\xi}) = \lambda\, f(t;\boldsymbol{\theta})$ *is the improper HR function related to the PTCR model and* $\mathrm{E}(\Upsilon\, S(t;\boldsymbol{\theta})^{\Upsilon-1}) = \sum_{j=1}^{\nu} j\,\tau_j\, S(t;\boldsymbol{\theta})^{j-1}$. *the equality in (ii) is attained only if*

$\tau_1 = 1$ and $\tau_2 = \tau_3 = \cdots = 0$, which corresponds to the PTCR model as we will show in the Example 2.1.

From (2), the cured fraction (p_0) is given by

$$p_0 = \lim_{t\to\infty} S_{\text{pop}}(t; \boldsymbol{\xi}) = \mathrm{e}^{-\lambda}. \tag{5}$$

Remark 2.2 The proposed model and the PTCR model have the same expression for the cured fraction.

The SF for the noncured population is given by

$$S^*(t; \boldsymbol{\xi}) = \Pr(T > t | D \geq 1) = \frac{\exp\left\{-\lambda\left[1 - \sum_{j=1}^{\nu} \tau_j \left(S(t; \boldsymbol{\theta})^j\right)\right]\right\} - \mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}}.$$

The PDF and HR functions related to the susceptible individuals are given by

$$f^*(t; \boldsymbol{\xi}) = \frac{f_{\text{pop}}(t; \boldsymbol{\xi})}{1 - \mathrm{e}^{-\lambda}}$$

and then

$$\begin{aligned} h^*(t; \boldsymbol{\xi}) &= \frac{f_{\text{pop}}(t; \boldsymbol{\xi})}{\exp\left\{-\lambda\left[1 - \sum_{j=1}^{\nu} \tau_j \left(S(t; \boldsymbol{\theta})^j\right)\right]\right\} - \mathrm{e}^{-\lambda}} \\ &= \left(\frac{\exp\left\{-\lambda\left[1 - \sum_{j=1}^{\nu} \tau_j \left(S(t; \boldsymbol{\theta})^j\right)\right]\right\}}{\exp\left\{-\lambda\left[1 - \sum_{j=1}^{\nu} \tau_j \left(S(t; \boldsymbol{\xi})^j\right)\right]\right\} - \mathrm{e}^{-\lambda}}\right) h_{\text{pop}}(t; \boldsymbol{\xi}) \\ &= \left(\frac{1}{\Pr(T < \infty | T > t)}\right) h_{\text{pop}}(t; \boldsymbol{\xi}), \end{aligned}$$

i.e., as expected, the hazard function is greater for the susceptible individuals than an individual selected from the complete population (Chen et al., 1999).

2.2 Special sub-models of the proposed model

There are more than 90 submodels in the literature within this class (Wimme and Altman 1996). We are motivated to introduce these new models because of the wide usage of Eq. (2) and the fact that the current generalization provides means of its extension to still more complex situations, with the hope that the new model may have a "better fit" in certain practical situations (see Sect. 4.2). However, we focused in some special models, including three that, to date, have not been proposed in the literature.

Example 2.1 (*Poisson cure rate model*). Consider $\nu = 1$ and $\tau_1 = 1$ in (1). The improper SF is given by

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = \exp\left\{-\lambda(1 - S(t;\boldsymbol{\theta}))\right\}, \quad t > 0.$$

This model is known in the literature as the PTCR model and was introduced by Yakovlev and Tsodikov (1996) and Chen et al. (1999). We use the notation $T \sim \text{PTCR}(\boldsymbol{\xi})$.

Example 2.2 (*Negative binomial cure rate model*). Consider $\nu = \infty$ and $\tau_i = -\frac{\tau^i}{i\log(1-\tau)}$, i.e., $\Upsilon_1, \Upsilon_2, \ldots,$ have logarithmic distribution. The improper SF is given by

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = \exp\left\{-\lambda\left[1 - \frac{\log\left(1 - \tau S(t;\boldsymbol{\theta})\right)}{\log(1-\tau)}\right]\right\}, \quad t > 0, \tag{6}$$

This model is named negative binomial cure rate model. We use the notation $T \sim \text{NBCR}(\boldsymbol{\xi})$. De Castro et al. (2009) introduced an alternative version of this model parameterized in the cure. The SF for such model is given by

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = \left[1 + (\exp(\lambda/\tau) - 1)\,(1 - S(t;\boldsymbol{\theta}))\right]^{-\tau}, \quad t > 0.$$

Such parametrization corresponds to take $\nu = \infty$ and $\tau_j = \frac{\tau[1-e^{-\lambda/\tau}]^j}{j}$. However, in this parametrization λ is included in the distribution of N and in the distribution of the $\Upsilon_j' s$. For interpretability, we prefer the parametrization in (6).

Example 2.3 (*Hermite cure rate model **(new!)***). Suppose now $\nu = 2$, $\tau_1 = 1/(1+\tau)$ and $\tau_2 = \tau/(1+\tau)$. In other words, $\Upsilon_1 - 1, \Upsilon_2 - 1, \ldots,$ have Bernoulli distribution with success probability $\tau/(1+\tau)$. The SF in this case is

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = \exp\left\{-\frac{\lambda}{1+\tau}\left(1 - S(t;\boldsymbol{\theta}) + \tau(1 - S(t;\boldsymbol{\theta})^2)\right)\right\}, \quad t > 0.$$

This model was not yet considered in the literature. We name this new model by Hermite cure rate model. We denoted as $T \sim \text{HERMCR}(\boldsymbol{\xi})$.

Example 2.4 (*Pólya cure rate model **(new!)***). Consider $\nu = \infty$ and $\tau_i = \tau\,(1-\tau)^{i-1}$, i.e., $\Upsilon_1, \Upsilon_2, \ldots,$ have geometric distribution. Thus,

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = \exp\left\{-\lambda\left[\frac{1 - S(t;\boldsymbol{\theta})}{1 - (1-\tau)S(t;\boldsymbol{\theta})}\right]\right\}, \quad t > 0.$$

This model is named Pólya cure rate model and was not yet considered in the literature. For $\tau = 1$, we deduce the PTCR model. We denoted as $T \sim \text{POLCR}(\boldsymbol{\xi})$.

Example 2.5 (*Thomas cure rate model* ***(new!)***). For the choice of $\nu = \infty$ and $\tau_i = \frac{\tau^{i-1}\exp(-\tau)}{(i-1)!}$. In other words, $\Upsilon_1, \Upsilon_2, \ldots,$ have a truncated Poisson distribution. In this case,

$$S_{\text{pop}}(t; \boldsymbol{\xi}) = \exp\left\{\lambda\left[S(t; \boldsymbol{\theta})\exp[-\tau\ (1 - S(t; \boldsymbol{\theta}))] - 1\right]\right\}.$$

This model is named Thomas cure rate model and was not yet considered in the literature. We denoted as $T \sim \text{THCR}(\boldsymbol{\xi})$.

Remark 2.3 The Weibull model is suitable for many biological problems. For this reason, we consider this model for the time-to-event of the carcinogenic cells with parametrization

$$S(t; \boldsymbol{\theta}) = \exp\left\{-\left(\frac{t\,\Gamma(1 + 1/\sigma)}{\mu}\right)^{\sigma}\right\},$$

where $\boldsymbol{\theta} = (\mu, \sigma)$, $\mathrm{E}(T) = \mu$ and $\mathrm{Var}(T) = \mu^2\left(\frac{\Gamma(2/\sigma+1)}{\Gamma^2(1/\sigma+1)} - 1\right)$. In the gamlss.dist R package, the function `dWEI3` define the density for this parametrization, where μ is the mean of the distribution (Stasinopoulos and Rigby 2007). Henceforth, we add the suffix WEI3 to the corresponding cure rate model to refers that we are using this model to the time-to-event for the concurrent causes. For instance, HERMCR-WEI3, POLCR-WEI3, etc.

3 Estimation based on a classical approach

In this section, we discuss the inference for the model based on a classical approach. Identifiability and computational aspects of the proposed model also is discussed.

3.1 The maximum likelihood estimators for the model

Note that the cure rate in (5) depends only on λ. In order to facilitate the introduction of covariates, from this moment we consider the parametrization in terms of $p_0 = e^{-\lambda}$. In a cure rate model framework, the individuals are subject to right censoring. Denote Y_i and C_i the failure and censoring times for the i-th individual, respectively, $i = 1, \ldots, m$. We observe $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \leq C_i)$, where $\delta_i = 1$ and $\delta_i = 0$ denote a failure and a censored time, respectively. We also assumed that is observed for each individual two sets of covariates, say $\mathbf{x}_i^\top$ and $\mathbf{z}_i^\top$ with dimensions r_1 and r_2 respectively, related to the time-to-event of the carcinogenic cells and the cure rate, respectively, and satisfying the following functional relations

$$g_1(\mu_i) = \eta_{1i} = \mathbf{x}_i^\top\boldsymbol{\beta} \quad \text{and} \quad g_2(p_{0i}) = \eta_{2i} = \mathbf{z}_i^\top\boldsymbol{\gamma}. \tag{7}$$

Furthermore, we assume that the covariate matrices $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^\top$ and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)^\top$ have ranks r_1 and r_2, respectively. The link functions $g_1 : \mathbb{R} \to \mathbb{R}^+$ and

$g_2 : \mathbb{R} \to (0, 1)$ must be strictly monotone, positive and at least twice differentiable, such that $\mu_i = g_1^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$ and $p_{0i} = g_2^{-1}(\mathbf{z}_i^\top \boldsymbol{\nu})$, with $g_1^{-1}(\cdot)$ and $g_2^{-1}(\cdot)$ being the inverse functions of $g_1(\cdot)$ and $g_2(\cdot)$, respectively. There are several possible choices for the link functions $g_1(\cdot)$ and $g_2(\cdot)$. For instance, two common specifications are the logarithmic function $g_1(\cdot) = \log(\cdot)$ and the logit function $g_2(\cdot) = \log(\cdot/(1-\cdot))$. Under this setup, the corresponding log-likelihood function for $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \tau)$ under non-informative censoring is expressed as

$$\ell(\boldsymbol{\xi}; \boldsymbol{D}_{obs}) = \sum_{i=1}^{m} \left[\delta_i \log f_{\text{pop}}(t_i; \boldsymbol{\xi}) + (1-\delta_i) \log S_{\text{pop}}(t_i; \boldsymbol{\xi})\right], \tag{8}$$

where $\boldsymbol{D}_{obs} = (\boldsymbol{t}, \boldsymbol{\delta}, \mathbf{X}, \mathbf{Z})$, with $\boldsymbol{t} = (t_1, \ldots, t_m)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ and f_{pop} and S_{pop} are presented in (3) and (2), respectively.

The maximum likelihood (ML) estimators $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\sigma}, \widehat{\tau})$ of $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \tau)$ are defined as the values of $\boldsymbol{\xi}$ that maximize the conditional log-likelihood function in (8). The ML estimators are obtained using numerical methods since equating the first-order log-likelihood derivatives to zero leads us to a complicated system of nonlinear equations. A future subsection discusses the computational implementation of the model.

3.2 Identifiability

Identifiability is an important property of any statistical model in order to guarantee basic conditions of the desirable estimators. In a cure rate models context, Li et al. (2001) discussed conditions in order to guarantee that the mixture model and the PTCR model be identifiable, and Hanin and Huang (2014) revisited such discussion in a more general framework for the cure rate models. Let $\boldsymbol{\xi} = (\boldsymbol{\beta}_1, \boldsymbol{\gamma}_1, \sigma_1, \tau_1)$ the vector of parameters. Also suppose that two (non-null) set of covariates $\mathbf{x} = (\mathbf{w}', \mathbf{x}')$ and $\mathbf{z} = (\mathbf{w}', \mathbf{z}')$ are available and introduced in μ and p_0 as $\mu_i = g_1(\mathbf{x}^\top \boldsymbol{\beta}_i)$ and $p_{0i} = g_2(\mathbf{z}^\top \boldsymbol{\gamma}_i)$, $i = 1, 2$, where $\mathbf{x}' \neq \mathbf{z}'$ and any of the three vectors ($\mathbf{x}'$, $\mathbf{z}$ and $\mathbf{w}'$) can be null. Note that $\mathbf{w}'$ is a common element for $\mathbf{x}$ and $\mathbf{z}$. This notation is very flexible because allows to denote different combinations for covariates $\mathbf{x}$ and $\mathbf{z}$. For instance, the case where both sets of covariates are equal ($\mathbf{x} = \mathbf{z}$) is represented for $\mathbf{x}'$ and $\mathbf{z}'$ being null; the case where both sets of covariates are different ($\mathbf{x} \neq \mathbf{z}$) is represented for $\mathbf{w}'$ being null and the case where no covariates for the susceptible part of the model are avaliable is represented for $\mathbf{w}'$ null and $\mathbf{x}' = \mathbf{1}_m$, a vector of ones of dimension m. The SF of the model can be written as

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = p_0(\boldsymbol{\gamma},\mathbf{z}) + (1 - p_0(\boldsymbol{\gamma},\mathbf{z})) \times \underbrace{\frac{1 - p_0(\boldsymbol{\gamma},\mathbf{z})^{-\sum\limits_{j=1}^{\nu} \tau_j S(t;\boldsymbol{\theta},\mathbf{x})^j}}{\left(1 - p_0(\boldsymbol{\gamma},\mathbf{z})^{-1}\right)}}_{S^*(t;\boldsymbol{\theta},\boldsymbol{\gamma},\tau_1,\tau_2,\ldots,\mathbf{x},\mathbf{z})},$$

i.e., as a mixture model with cure rate $p_0(\boldsymbol{\gamma},\mathbf{z})$ and survival function for the time-to-event in the susceptible individuals $S^*(t;\boldsymbol{\theta},\boldsymbol{\gamma},\tau_1,\tau_2,\ldots,\mathbf{x},\mathbf{z})$. Note that by point 2 in Theorem 1 of Li et al. (2001) the model is not identifiable if $p_0(\boldsymbol{\gamma},\mathbf{z})$ is constant (i.e., if $\mathbf{z}$ includes only the term related to the intercept). Therefore, the first and basic condition to make the model identifiable is that $\mathbf{z}$ includes at least one covariate. Let

$$\begin{aligned} ||p_0||_{\mathbf{w}'} &= \sup\{p_0(\boldsymbol{\gamma},\mathbf{w}',\mathbf{z}') : \mathbf{z}' \in \mathcal{Z}\} \quad \text{and} \\ ||F^*||_{\mathbf{w}'} &= \sup\{\lim_{t\to\infty} F^*(t;\boldsymbol{\theta},\boldsymbol{\gamma},\tau_1,\tau_2,\ldots,\mathbf{w}',\mathbf{x}',\mathbf{z}') : \mathbf{x}' \in \mathcal{X}\}. \end{aligned}$$

As $p_0(\boldsymbol{\gamma},\mathbf{z}) = g_2(\mathbf{z}^\top\boldsymbol{\gamma}_i)$, $g_2 : \mathbb{R} \to (0,1)$ and is monotone, follows that $||p_0||_{\mathbf{w}'} = 1$. On the other hand, as $0 \le \tau_1,\tau_2,\ldots,S(t;\boldsymbol{\theta},\mathbf{x}) \le 1$, then it is clear that $0 \le \sum_{j=1}^{\nu}\tau_j S(t;\boldsymbol{\theta},\mathbf{x})^j \le 1$, $\forall p_1, p_2, \ldots$, $\forall\boldsymbol{\theta} \in \Theta, \mathbf{x} \in \mathcal{X}$. As we are assuming that $S(\cdot;\boldsymbol{\theta})$ is a proper function, follows that $F^*(t;\boldsymbol{\theta},\boldsymbol{\gamma},\tau_1,\tau_2,\ldots,\mathbf{w}',\mathbf{x}',\mathbf{z}') \in [0,1]$ and then $||F^*||_{\mathbf{w}'} = 1$. By point 1 in Theorem 1 of Hanin and Huang (2014), the model is identifiable.

Remark 3.1 Note that the unique condition to the model be identifiable is that $\mathbf{z}$ includes at least one covariate. This allows us to make the following statements

(i) The model is identifiable if no covariates are considered to model the mean of the promotion times ($\mathbf{w}'$ is null and $\mathbf{x} = \mathbf{1}_m$).
(ii) The model is identifiable if the same covariates are considered to model the cure rate and the promotion times ($\mathbf{x}'$ and $\mathbf{z}'$ are null).

Remark 3.2 For the HERMCR-WEI3 model we have that $\boldsymbol{\xi} = (\lambda,\tau,\mu,\sigma)$, with SF given by

$$S_{\text{pop}}(t;\boldsymbol{\xi}) = \exp\left\{-\frac{\lambda}{1+\tau}\left(1 - \exp\left\{-(y\zeta)^{\sigma}\right\} + \tau\left(1 - \exp\left\{-2(y\zeta)^{\sigma}\right\}\right)\right)\right\},$$

where $\zeta = \Gamma(1+1/\sigma)/\mu$. Note that $\boldsymbol{\xi}_1 = (\lambda,0,\mu,\sigma)$ and $\boldsymbol{\xi}_2 = (\lambda,\infty,2^{-1/\sigma}\mu,\sigma)$ satisfies that $\boldsymbol{\xi}_1 \neq \boldsymbol{\xi}_2$ and $S_{\text{pop}}(t;\boldsymbol{\xi}_1) = S_{\text{pop}}(t;\boldsymbol{\xi}_2)$, $\forall t > 0$. In other word, this model is not identifiable. This problem is similar to the obtained in the proportional hazards model when an intercept is included. For this reason, for this model we also necessarily considered covariates in μ and without intercept term.
The same problem is identified for any distribution where $S(t;\boldsymbol{\theta})$ and $[S(t;\boldsymbol{\theta})]^2$ belongs to the same class of models. For instance, any distribution in the Lehman type II family of distributions (Gupta et al. 1998) with positive support satisfies this condition.

3.3 Computational aspects

The models presented in Sect. 2.2 were implemented in the `gamlss` framework (Rigby and Stasinopoulos 2005) to facilitate their use by researchers from other areas. For instance, for the HERM/WEI3 model with two covariates in the cure rate and the mean of the concurrent causes can be fitted as

```
gamlss(Surv(y, delta) ~ x1+x2, family = cens(HERMWEI3),
               nu.formula = ~x1+x2)
```

The model also was implemented maximizing directly the log-likelihood function in (8) using the `nlminb` function. The estimated Hessian matrix, say $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\xi})$, was estimated based on a numerical approximation given by the `pracma` package. Since $\widehat{\boldsymbol{\xi}}$ is a maximum likelihood estimator of $\boldsymbol{\xi}$, under suitable regularity conditions, it can be shown that (see Kalbfleisch and Prentice 2002, page 60)

$$\sqrt{m}\left[\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\xi})\right]^{-1/2}\left(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}\right)\overset{\mathcal{D}}{\to} N_{r_1+r_2+2}(\mathbf{0}_{r_1+r_2+2},\mathbf{I}_{r_1+r_2+2}), \qquad \text{as } m\to\infty,$$

where $\mathbf{0}_r$ and $\mathbf{I}_r$ denote a vector of zeros of dimension r and the identity matrix of order r, respectively.

4 Numerical examples

In this section, we present a Monte Carlo simulation study and the real data problem related to melanoma cancer in the state of São Paulo discussed in the introduction section.

4.1 A simulation study

We present a simulation study in order to assess the properties of the ML estimators in finite samples for our proposal. We considered the HERMCR-WEI3, POLCR-WEI3 and THCR-WEI3 models. We considered two covariates (say z_1 and z_2) related to the cure rate and two covariates (x_1 and x_2) related to the promotion times of the cells. z_1 and x_1 were drawn independently from the uniform model between 0 and 2 and z_2 and x_2 were drawn independently from the Bernoulli distribution with success probabilities 0.4 and 0.7, respectively. In all the cases, the true parameters were fixed as $\beta_0 = 1.0$, $\beta_1 = -0.8$, $\beta_2 = -0.5$, $\gamma_0 = 1.5$, $\gamma_1 = -0.6$, $\gamma_2 = -0.7$ and $\log(\sigma) = 1.1$. For the HERMCR-WEI3 and THCR-WEI3 models, we also considered $\log(\tau) = \log(0.35) \approx -1.05$ and for the POLCR-WEI3 model we considered $\text{logit}(\tau) \approx -0.62$, were $\text{logit}(x) = \log(x/(1-x))$. Such covariates were introduced as

$$\begin{aligned}\log(\mu_i) &= \eta_{1i} = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} \qquad \text{and}\\ \text{logit}(p_{0i}) &= \eta_{2i} = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}, \quad i = 1\ldots, m.\end{aligned}$$

To avoid the identifiability problems of the HERMCR-WEI3 model, we don't consider the intercept term in the promotion times (i.e., $\beta_0 = 0$ in this case). We considered four sample sizes: 200, 300, 500 and 1000. For each combination of model and sample size, we drawn the covariates and then they were fixed through the 1000 replicates that we considered. To draw values from the model and for $i = 1, \ldots, m$, we simulate $N_i \sim \text{Po}(\theta_i)$. If $N_i > 0$, we draw $X_1, \ldots, X_{N_i}$ from the respective model (Bernoulli, geometric or truncated Poisson for HERMCR-WEI3, POLCR-WEI3 and THCR-WEI3, respectively) and define $D_i = X_1 + \ldots + X_{N_i}$. Therefore, we draw $V_1, \ldots, V_{D_i}$ from the WEI3 model in Remark 2.3. Finally, the failure times are defined as $T_i^* = \min(V_1, \ldots, V_{D_i})$, for $N_i > 0$ and $T_i^* = \infty$, for $N_i = 0$. To incorporate a censoring scheme, we consider $C_i = 20$, i.e., a type I censoring scheme. Finally, the observed times are defined as $T_i = \min(T_i^*, C_i)$ and the failure indicators are $\delta_i = I(T_i^* \leq C_i)$. For each sample, we apply the ML estimation in R (R Core Team 2020) considering: i) the `gamlss` framework (say $\widehat{\boldsymbol{\xi}}_1$); and ii) the direct maximization of the log-likelihood function with the `nlminb` function (say $\widehat{\boldsymbol{\xi}}_2$). The estimated vector of parameters was considered as $\widehat{\boldsymbol{\xi}}$, where $\ell(\widehat{\boldsymbol{\xi}}; \boldsymbol{D}_{obs}) = \max(\ell(\widehat{\boldsymbol{\xi}}_1; \boldsymbol{D}_{obs}), \ell(\widehat{\boldsymbol{\xi}}_2; \boldsymbol{D}_{obs}))$, i.e., the point where the log-likelihood function in (8) attaches the maximum between the two estimated points. We also computed the estimated standard error based on the Hessian matrix. For each parameter, the results are summarized with the mean of the estimated bias, the root of the estimated mean squared errors (RMSE), the mean of the estimated standard errors (SE) and the coverage probabilities (CP) of the asymptotic 95% confidence intervals. In all the cases, the average of the percentage of censored observations was around 65%.

Table 1 presents the biases, RMSE and CP of the estimators of the parameters (using the maximum likelihood estimation) for the Hermite, Pólya and Thomas models. As expected, increasing the sample size reduces substantially the RMSE in all cases considered. We also observe that the MLEs of β_0, β_1, β_2, γ_0, γ_1, γ_2 and $\log(\sigma)$ work well for all cases considered. On the other hand, we call attention for the MLE of $\text{logit}(\tau)$. The $\text{logit}(\tau)$ parameter presented higher bias and RMSE. Therefore, a larger sample size is necessary to obtain satisfactory results with respect to the estimation of $\text{logit}(\tau)$ for the HERMCR-WEI3, POLCR-WEI3 and THCR-WEI3 models. Furthermore, note that the asymptotic confidence intervals have an empirical coverage probability that is less than the nominal value 0.95, but the coverage probability for $\text{logit}(\tau)$ is higher than the nominal level for small sample size. Overall, we observe that the asymptotic confidence intervals have a good performance.

4.2 Melanoma cancer data

The melanoma dataset comprises 6749 records of patients diagnosed with melanoma in the state of São Paulo, Brazil, between 2000 and 2014, with follow-up conducted until 2018. All records were provided by the São Paulo Oncocenter Foundation (FOSP), and it can be downloaded in http://www.fosp.saude.sp.gov.br. The hospital cancer registry (RHC/FOSP) started its activities in 2000, intending to register cancer cases treated in the state. Currently, 77 hospital cancer registries are active, and every three months, the records send the datasets. The FOSP is a public institution connected to the State Health Secretariat, which assists in preparing and implementing healthcare policies in Oncology. As mentioned by Andrade et al. (2012), these policies serve

Table 1 Estimated bias (bias), root of the estimated mean squared error (RMSE), mean of the estimated standard errors (se) and approximated 95% coverage probabilities (CP)

Model	Parameter	$m = 200$				$m = 300$				$m = 500$				$m = 1000$			
		Bias	RMSE	SE	CP	Bias	RMSE	SE	CP	Bias	RMSE	SE	CP	Bias	RMSE	SE	CP
HERMCR-WEI3	β_1	− 0.0001	0.0446	0.0426	0.933	− 0.0009	0.0444	0.0445	0.941	− 0.0002	0.0296	0.0290	0.936	− 0.0027	0.0209	0.0206	0.944
	β_2	− 0.0086	0.0920	0.0927	0.942	− 0.0058	0.0782	0.0778	0.952	− 0.0001	0.0586	0.0589	0.947	0.0022	0.0417	0.0419	0.945
	$\log(\sigma)$	0.0230	0.0989	0.0931	0.932	0.0123	0.0763	0.0752	0.943	0.0063	0.0561	0.0548	0.941	0.0054	0.0409	0.0405	0.950
	γ_0	0.0221	0.3173	0.3051	0.944	0.0263	0.2825	0.2739	0.946	0.0153	0.1996	0.1980	0.956	0.0068	0.1381	0.1387	0.958
	γ_1	− 0.0146	0.1911	0.1837	0.943	− 0.0177	0.1491	0.1440	0.951	− 0.0129	0.1113	0.1137	0.959	− 0.0023	0.0786	0.0785	0.941
	γ_2	− 0.0145	0.3666	0.3551	0.945	− 0.0237	0.3184	0.3093	0.951	− 0.0020	0.2174	0.2216	0.951	− 0.0041	0.1556	0.1594	0.962
	$\log(\tau)$	0.7850	4.4125	29.240	1.000	0.3036	2.4737	8.3866	1.000	0.1524	1.3573	1.7304	1.000	0.0365	0.5028	0.4927	0.986
POLCR-WEI3	β_0	0.0453	0.1627	0.1842	0.966	0.0280	0.1308	0.1480	0.949	0.0194	0.1064	0.1196	0.954	0.0054	0.0780	0.0866	0.958
	β_1	0.0007	0.0737	0.0704	0.935	0.0030	0.0571	0.0553	0.938	− 0.0030	0.0439	0.0441	0.950	0.0003	0.0310	0.0307	0.944
	β_2	− 0.0010	0.0849	0.0829	0.939	− 0.0036	0.0674	0.0667	0.943	0.0010	0.0552	0.0512	0.929	0.0004	0.0352	0.0364	0.952
	$\log(\sigma)$	0.0583	0.1294	0.1366	0.906	0.0392	0.1076	0.1141	0.911	0.0249	0.0863	0.0935	0.930	0.0092	0.0635	0.0687	0.942
	γ_0	0.0228	0.3620	0.3613	0.955	0.0022	0.3213	0.3142	0.955	0.0055	0.2415	0.2350	0.944	0.0092	0.1683	0.1698	0.952
	γ_1	− 0.0076	0.2534	0.2522	0.947	− 0.0081	0.1956	0.1960	0.955	− 0.0072	0.1641	0.1621	0.945	− 0.0077	0.1188	0.1170	0.947
	γ_2	− 0.0133	0.3309	0.3212	0.949	0.0097	0.2800	0.2733	0.940	0.0009	0.2003	0.2018	0.949	− 0.0026	0.1401	0.1429	0.952
	$\text{logit}(\tau)$	1.3193	5.4474	76.059	0.941	1.1082	4.7100	44.059	0.935	0.8410	3.9130	22.675	0.939	0.3962	2.3509	6.5390	0.943
THCR-WEI3	β_0	− 0.0505	0.1734	0.2052	0.924	− 0.0304	0.1545	0.1857	0.929	− 0.0101	0.1482	0.1623	0.937	− 0.0067	0.1389	0.1280	0.931
	β_1	0.0028	0.0736	0.0707	0.945	− 0.0036	0.0574	0.0565	0.937	0.0014	0.0457	0.0446	0.942	0.0019	0.0324	0.0315	0.948
	β_2	− 0.0045	0.0831	0.0834	0.938	0.0007	0.0678	0.0684	0.956	0.0020	0.0545	0.0526	0.941	0.0000	0.0371	0.0377	0.948
	$\log(\sigma)$	0.0040	0.0794	0.0935	0.966	− 0.0032	0.0713	0.0777	0.951	− 0.0073	0.0540	0.0602	0.959	− 0.0099	0.0417	0.0438	0.965
	γ_0	0.0427	0.3813	0.3625	0.941	0.0145	0.3357	0.3220	0.940	0.0154	0.2341	0.2379	0.958	−0.0032	0.1589	0.1615	0.965
	γ_1	− 0.0284	0.2532	0.2529	0.952	− 0.0061	0.2132	0.2066	0.944	− 0.0108	0.1607	0.1596	0.959	0.0014	0.1087	0.1107	0.959
	γ_2	− 0.0194	0.3312	0.3219	0.946	− 0.0077	0.2792	0.2750	0.955	− 0.0081	0.1964	0.2006	0.957	− 0.0050	0.1390	0.1396	0.948
	$\log(\tau)$	− 1.3941	4.2154	43.289	0.997	− 1.1123	3.7570	25.799	0.998	− 0.5142	2.3965	7.8060	0.993	− 0.2283	1.0847	1.5663	0.970

as an instrument for oncology hospitals to prepare their protocols and improve care practices.

In the routine clinical the staging system proposed by the American Joint Committee on Cancer (AJCC) is used to define stage melanoma cases. As mentioned by Calsavara et al. (2020), the early clinical stages (I or II) are associated with a better prognosis. The great majority of patients with stage I or II melanoma will be alive after 10 years of follow-up, once that most of these cases are treated with surgery. A worst prognosis is expected in patients diagnosed in the clinical stage III or IV, and the specific survival rates at 10 years after diagnosis may vary from 24 to 88% (Gershenwald et al. 2017). Here, we analyzed the effect of all covariates such as age at diagnosis (in years), sex, surgery, clinical stage, radiotherapy, and chemotherapy.

In our study, the event of interest was defined as death due to cancer. The main goals were to identify the effects of the observed covariates, including sex, age at diagnosis, clinical stage, surgery, radiotherapy and chemotherapy, on the mean of the time-to-event, as well as in the long-term survivors.

We fitted the discussed models in Sect. 2.2 to this data set considering the covariates `surgery`, `clinical stage`, `age`, `sex`, `radiotherapy` and `chemotherapy` in the cure rate (with the `logit` link) and the mean of the carcinogenic cells (with the `log` link). The results of AIC and BIC criteria for each fitted model are shown in Table 2. According to the AIC and BIC criteria, the POLCR-WEI3 model seems to be better choice among the fitted models and it will be our working model, thus we will focus exclusively on the interpretation of POLCR-WEI3 model parameters.

Table 3 shows the estimated parameters for the selected model. According to results, among the observed covariates considered in the model, there is evidence that all variables, except surgery, are important factors to explain the long-term survivors (component $\boldsymbol{\gamma}$). As expected, as clinical stage increases the cure rate decreases. Age at diagnosis is associated with the cure rate, as age increases the cure rate decreases. Lower cure rates are expected for patients who received radiotherapy and chemotherapy. Regarding to the effect of covariates in the mean of the concurrent causes (component $\boldsymbol{\beta}$) the surgery, clinical stage, radiotherapy, and chemotherapy were statistically significant. According to results, higher mean of the concurrent causes are expected to surgery, radiotherapy, and chemotherapy covariates ($\widehat{\beta} > 0$), while lower mean is expected for clinical stage advanced. Note that the best prognosis is given for patients in stage I, receiving surgery, female and without radiotherapy and chemotherapy, while the worst prognosis is given for patients in stage IV, without surgery, male with radiotherapy and chemotherapy. Based on such profiles (assuming a reasonable age limit of 100 years), the cure rate vary from 0.004 to 0.960. Then, using the relation $\lambda = -\log p_0$, we obtain that, for this particular problem, the mean of the number of

Table 2 AIC and BIC criteria for the fitted models in melanoma data set

Model	PTCR-WEI3	HERMCR-WEI3	NBCR-WEI3	POLCR-WEI3	THCR-WEI3
AIC	10931.1	10914.6	10886.2	**10866.1**	10899.6
BIC	11060.6	11050.9	11022.5	**11002.4**	11035.9

Bold values indicate the lowest AIC and BIC values

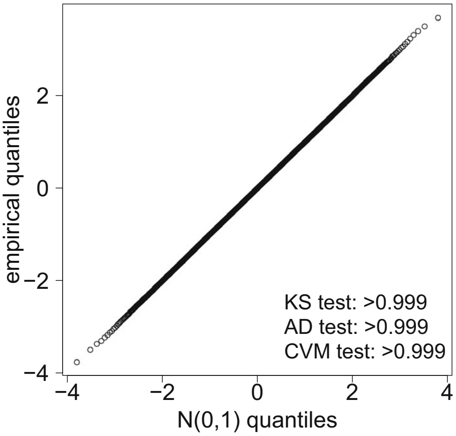
Table 3 Maximum likelihood estimate, standard error (SE), z-value and p value for POLCR-WEI3 considering sex, age, clinical stage, surgery, radiotherapy and chemotherapy for the melanoma cancer dataset

Component	Covariate	Estimate	SE	z-value	p value
$\boldsymbol{\beta}$	Intercept	2.9588	0.6033	4.9042	< 0.0001
	surgery	0.6083	0.0861	7.0623	< 0.0001
	stage II	− 0.0396	0.1524	− 0.2595	0.7952
	stage III	− 0.5988	0.1392	− 4.3010	< 0.0001
	stage IV	− 1.4024	0.1419	− 9.8851	< 0.0001
	age	0.0025	0.0022	1.1341	0.2568
	sex: Female	0.0761	0.0702	1.0835	0.2786
	radiotherapy	0.3394	0.0990	3.4300	0.0006
	chemotherapy	0.6646	0.0847	7.8473	< 0.0001
$\boldsymbol{\gamma}$	Intercept	2.4104	0.2743	8.7863	< 0.0001
	surgery	0.2925	0.1620	1.8061	0.0709
	stage II	− 1.3986	0.1697	− 8.2414	< 0.0001
	stage III	− 1.9403	0.1577	− 12.3019	< 0.0001
	stage IV	− 2.9037	0.1922	− 15.1072	< 0.0001
	age	− 0.0203	0.0034	− 5.9013	< 0.0001
	sex: Female	0.4654	0.1087	4.2811	< 0.0001
	radiotherapy	− 1.4406	0.2646	− 5.4446	< 0.0001
	chemotherapy	− 1.4948	0.1888	− 7.9169	< 0.0001
$\log(\sigma)$	–	0.3879	0.0260	14.9375	< 0.0001
$\text{logit}(\tau)$	–	− 2.9470	0.8840	− 3.3337	0.0009

clusters range from 0.041 to 5.463. Therefore, considering the 99.9% percentile of the Poisson distribution, we conclude that for the better prognosis at most there is a cluster of cells and for the worse prognosis at most there is 12 clusters of cells. On the other hand, by the biological motivation for the POLCR-WEI3 model given in Sect. 2, it is deduced that the distribution for the number of carcinogenic cells in each cluster follows a geometric distribution with parameter $\widehat{\tau} = 0.05$. This implies that the estimated mean and standard deviation for the carcinogenic cells in each cluster are 20.05 and 19.54, respectively.

In order to justify the inclusion of covariates in the mean of the concurrent causes, we considered that $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta}^{\top} = (\beta_{\texttt{Intercept}}, \boldsymbol{\beta}^{\star\top})$, i.e., $\boldsymbol{\beta}^{\star}$ is the vector of parameters related to the covariates in this part of the model. Therefore, we perform the test $H_0 : \boldsymbol{\beta}^{\star} = \mathbf{0}_8$ versus $H_1 : \boldsymbol{\beta}^{\star} \neq \mathbf{0}_8$. The statistic for the likelihood ratio test is given by $\text{LR} = 2(-5413.04 - (-5568.32)) = 310.57$ with an associated p value < 0.0001 (based on the chi-squared distribution with 7 freedom degrees). Then, the inclusion of covariates in the mean of the cells is widely supported. Figure 3 shows the qq-plot for the randomized quantile residuals (RQRs) (Dunn and Smyth 1996) for the selected model. If the model were correctly specified for the data, the RQRs represent a random sample from the standard normal model, which is supported by

Fig. 3 Randomized quantile residuals for the POLCR-WEI3 model in melanoma data set



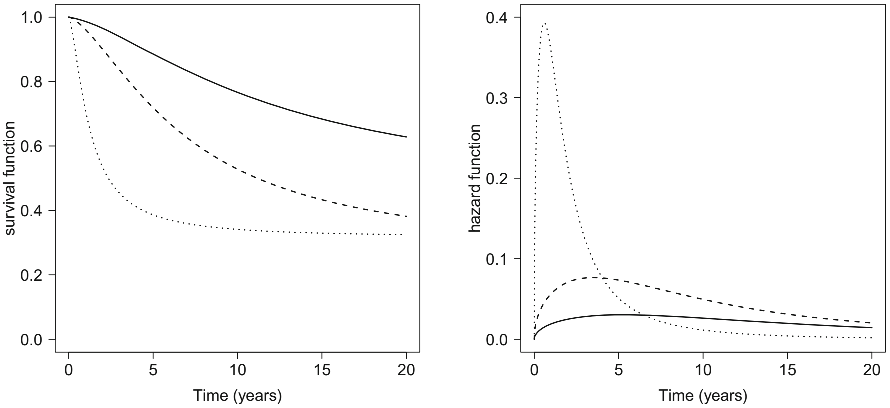


Fig. 4 Estimated survival function (left) and hazard function (right) obtained for the POLCR-WEI3 model for different profiles: (1) with surgery, clinical stage I, 20 years old, female, with radiotherapy and chemotherapy (solid curve); (2) with surgery, clinical stage II, 40 years old, male, with chemotherapy but not radiotherapy (dashed curve) and; (3) without surgery, clinical stage III, 60 years old, male, without radiotherapy or radiotherapy (dotted curve)

the Kolmogorov–Smirnov (KS), Anderson–Darling (AD) and the Cramér–Von-Mises (CVM) test. Therefore, we conclude that the model is appropriate for this data set.

Figure 4 shows the estimated survival function and hazard function for three selected profiles. As expected, younger female patients with early-stage cancer (clinical stage I), who had undergone surgery and who received radiotherapy and chemotherapy have higher survival; whereas 60 years-old male patients, diagnosed in clinical stage III, without surgery, who did not receive radiotherapy or chemotherapy had a worse survival function.

5 Concluding remarks

There have been significant recent works-both practical and theoretical-focusing on the use of the promotion cure rate model studied in Yakovlev and Tsodikov (1996) and Chen et al. (1999). Motivated from a new biological interpretation of cancer metastasis, in this paper, we introduced a general method for obtaining more flexible cure rate models and extended the promotion cure rate model. We have assumed a compound Poisson distribution for the number of cells, a Poisson distribution for the number of clusters of cells for an individual left active after the initial treatment, and some discrete distribution (truncated at zero) for the number of cells in jth cluster. Some mathematical properties of the new cure rate model was studied. Maximum likelihood inference is implemented straightforwardly for estimating the model parameters. We then conducted a simulation study to establish their empirical properties in order to evaluate their performances. In the empirical application, the proposed cure rate models show the potential of using the new methodology. In conclusion, we define a general approach for generating new cure rate models, at least 90 models (Wimme and Altman 1996), some of them known and the great majority new ones. The practical relevance and applicability of the proposed models were demonstrated using a real dataset of patients diagnosed with melanoma. As expected, the observed covariates: age at diagnosis, sex, clinical stage, radiotherapy, and chemotherapy covariates were important factors to explain the long-term survivors, whereas the covariates surgery, clinical stage, radiotherapy, and chemotherapy were statistically significant to explain the mean of the concurrent causes. Further, we motivate the use of the new cure rate model from a new biological interpretation of cancer metastasis. We think these two facts combined may attract more complex applications in the literature of survival analysis. Future work should explore other estimation methods for the proposed cure rate model, for instance, the Bayesian approach similarly as developed by Chen et al. (1999).

References

Berrino E et al (2021) High BRAF variant allele frequencies are associated with distinct pathological features and responsiveness to target therapy in melanoma patients. ESMO Open 6:100133

Calsavara VF, Milani EA, Bertolli E, Tomazella V (2020) Long-term frailty modeling using a non-proportional hazards model: application with a melanoma dataset. Stat Methods Med Res 29:2100–2118

Chen T, Du P (2018) Promotion time cure rate model with nonparametric form of covariate effects. Stat Med 37:1625–1635

Chen M-H, Ibrahim JG, Sinha D (1999) A new Bayesian model for survival data with a surviving fraction. J Am Stat Assoc 94:909–919

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

Coordenação de Prevenção e Vigilância (2017). Instituto Nacional de Cancer José Alencar Gomes da Silva. Estimativa 2018: Incidência de Cancer no Brasil. Coordenação de Prevenção e Vigilância. Rio de Janeiro. http://www1.inca.gov.br/estimativa/2018/

de Andrade CT, Magedanz AMPCB, Escobosa DM, Tomaz WM, Santinho CS, Lopes TO, Lombardo V (2012) The importance of a database in the management of healthcare services. Einstein (São Paulo) 10:360–365

De Castro M, Cancho VG, Rodrigues J (2009) A Bayesian long-term survival model parametrized in the cured fraction. Biom J 51:443–455

Dunn P, Smyth G (1996) Randomized quantile residuals. J Comput Graph Stat 5:236–244

Ervik M, Lam F, Ferlay J, Mery L, Soerjomataram I, Bray F et al (2016) Cancer today Lyon, France: international agency for research on cancer. 2016. Cancer today. https://www.gco.iarc.fr/today. Accessed 01/02/2019

Gershenwald JE, Scolyer RA, Hess KR, Sondak VK, Long GV, Ross MI, Lazar AJ, Faries MB, Kirkwood JM, McArthur GA et al (2017) Melanoma staging: evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin 67:472–492

Gómez YM, Gallardo DI, Leão J, Calsavara VF (2021) On a new piecewise regression model with cure rate: diagnostics and application to medical data. Stat Med 40(29):6723–6742

Gupta RC, Gupta PL, Gupta RD (1998) Modeling failure time data by Lehman alternatives. Commun Stat Theory Methods 27:887–904

Hanin L, Huang L (2014) Identifiability of cure models revisited. J Multivar Anal 130(1):261–274

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. Wiley, New York

Kim S, Chen MH, Dey DK (2011) A new threshold regression model for survival data with a cure fraction. Lifetime Data Anal 17:101–122

Leão J, Bourguignon M, Saulo H, Santos-Neto M, Calsavara V (2021) The negative binomial beta prime regression model with cure rate: application with a melanoma dataset. J Stat Theory Pract 15(3):1–21

Li CS, Taylor JM, Sy JP (2001) Identifiability of cure models. Stat Probab Lett 54(4):389–395

Molina KC, Calsavara VF, Tomazella VD, Milani EA (2021) Survival models induced by zero-modified power series discrete frailty: application with a melanoma data set. Stat Methods Med Res 29(8):2100–2118

Puglisi R et al (2021) Biomarkers for diagnosis, prognosis and response to immunotherapy in melanoma. Cancers (Basel) 13:2875

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape, (with discussion). Appl Stat 54(3):507–554

Rodrigues J, de Castro M, Cancho V, Balakrishnan N (2009) COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. J Stat Plan Inference 139:3605–3611

Rodrigues J, de Castro M, Balakrishnan N (2011) Destructive weighted Poisson cure rate models. Lifetime Data Anal 2011(17):333–346

Rodrigues J, Cordeiro GM, Cancho VG, Balakrishnan N (2016) Relaxed Poisson cure rate models. Biom J 58:397–415

Rodrigues AS, Calsavara VF, Bertolli E, Peres SV, Tomazella VL (2021) Bayesian long-term survival model including a frailty term: application to melanoma data. Chil J Stat 12(1):53–70

Shain AH, Bastian BC (2016) From melanocytes to melanomas. Nat Rev Cancer 16:345–358

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

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

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

Tucker S, Thames H, Taylor J (1990) How well is the probability of tumor cure after fractionated irradiation described by Poisson statistics? Radiat Res 24:273–282

Wimme G, Altman G (1996) The multiple Poisson distribution, its characteristics and a variety of forms. Biom J 38:995–1011

Yakovlev AY, Tsodikov AD (1996) Stochastic model of tumor latency and their biostatistical applications. World Scientific, Singapore

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.