Received: 25 September 2019 | Revised: 15 May 2020 | Accepted: 26 May 2020

DOI: 10.1002/sim.8664

RESEARCH ARTICLE

Statistics in Medicine WILEY

A new cure rate model with flexible competing causes with applications to melanoma and transplantation data

Jeremias Leão[1] | **Marcelo Bourguignon**[2] | **Diego I. Gallardo**[3] | **Ricardo Rocha**[4] | **Vera Tomazella**[5]

[1]Department of Statistics, Universidade Federal do Amazonas, Amazonas, Brazil

[2]Department of Statistics, Universidade Federal do Rio Grande do Norte, Natal, Brazil

[3]Department of Mathematics, Universidad de Atacama, Copiapó, Chile

[4]Department of Statistics, Universidade Federal da Bahia, Salvador, Brazil

[5]Department of Statistics, Universidade Federal de São Carlos, São Paulo, Brazil

Correspondence
Marcelo Bourguignon, Department of Statistics, Universidade Federal do Rio Grande do Norte, Natal, Brazil.
Email: m.p.bourguignon@gmail.com

In this article, we introduce a long-term survival model in which the number of competing causes of the event of interest follows the zero-modified geometric (ZMG) distribution. Such distribution accommodates equidispersion, underdispersion, and overdispersion and captures deflation or inflation of zeros in the number of lesions or initiated cells after the treatment. The ZMG distribution is also an appropriate alternative for modeling clustered samples when the number of competing causes of the event of interest consists of two subpopulations, one containing only zeros (cure proportion), while in the other (noncure proportion) the number of competing causes of the event of interest follows a geometric distribution. The advantage of this assumption is that we can measure the cure proportion in the initiated cells. Furthermore, the proposed model can yield greater or lower cure proportion than that of the geometric distribution when modeling the number of competing causes. In this article, we present some statistical properties of the proposed model and use the maximum likelihood method to estimate the model parameters. We also conduct a Monte Carlo simulation study to evaluate the performance of the estimators. We present and discuss two applications using real-world medical data to assess the practical usefulness of the proposed model.

KEYWORDS

cure rate models, long-term survival model, medical data, Weibull distribution, zero-modified geometric distribution

1 | INTRODUCTION

In medical and epidemiological studies, often interest focuses on studying the effect of concomitant information on the time to an event, such as death or recurrence of a disease. When the primary interest is to estimate the covariate effect, the Cox proportional hazards model is commonly used in the analysis of survival time data.[1] With the development of medical and health sciences, the data sets collected from clinical studies pose some new challenges to statisticians. New statistical models that can incorporate these changes should be investigated. The most prevalent change noted in several clinical studies is that more patients respond favorably to treatment or they were not susceptible to the event of interest in the study, so they are considered cured or have prolonged disease-free survival. The modeling of the cure fraction (proportion of cured patients) has been widely developed in the biostatistics literature. Historically, one of the most famous cure rate models is the mixture cure model introduced by Berkson and Gage.[2] This model has been

extensively discussed by several authors, including References 3-7. Further development for cure rate models was obtained in two ways:

1. In a competing risk scenario. Initial works in this line are References 8 and 9, where the promotion time cure model or bounded cumulative hazard model in cancer relapse settings was proposed, assuming that a latent biological process of N latent carcinogenic tumor cells generates the observed failure (relapse), which is given by the minimum time related to each of the N tumor cells. In such a proposal, the authors considered the Poisson distribution to model N. We remark that, in this context, $N = 0$ is defined as a "cured individual." Later, different proposals have been considered for N in the literature, namely, negative binomial and geometric,[10] COM-Poisson,[11] power series,[12] and Yule-Simon.[13] Other extensions in a cure rate model framework consider a more general activation scheme (not only the minimum, see References 14-16 for details) and a destructive scheme (ie, considering that each of the N initial cells has a probability p of being "activated"[17,18]). To avoid identifiability problems under this setting, a usual assumption is that the cure term is modeled by covariates (see References 19,20 for details).
2. In a frailty model context. Studies on frailty models generally assume a nonnegative and continuous frailty random variable. However, a continuous frailty distribution does not allow the existence of zero risks. In a cure rate model context, frailty "zero" can be interpreted as the existence of a subgroup of nonsusceptible individuals where the event of interest has not happened even after a long period of observation. In this context,[21,22] emphasized the difficulty of obtaining models in which the case of homogeneous frailty is a natural special case of the general model in which frailty varies randomly across units Let W be a continuous random variable following a frailty distribution, then to obtain identifiability we need $E(W) < \infty$, where it is often convenient to impose $E(W) = 1$ by a suitable constraint on the parameters of the frailty distribution. The same identifiability problem does not arise with discrete frailty because the frailty distribution lacks a scale parameter. However, this is also a disadvantage as the same approach of fixing $E(W)$ and allowing $\text{Var}(W) = 1$ to govern the strength of the frailty cannot be applied. References 23,24 discuss some strategies to accommodate zero frailty, for example, if $W = 0$ is used to describe long-term survivors, then units will never fail. Other examples of this approach are given in References 12,25-28.

A third and recent way to introduce cure rate models is known in the literature as the improper models introduced by Balka et al.[29] In these models, the parameter space of the distribution is extended producing an improper model in which the cumulative distribution function no longer approaches 1, but to $p \in (0, 1)$. More details in improper models extensions can be found in References 30-34. However, differently from the competing risks or frailty approaches, improper models do not have an underlying motivation and further connections with real problems are more complicated. We highlight that the first scheme is more flexible than the one that imposes the restriction $E(W) = 1$ (the second scheme), despite in some cases, where covariates are not available, the second approach is implied. As mentioned, under the competing risks scheme, the cure depends on $P(N = 0)$, which in turn depends on the available covariates. However, in certain problems, there could be relevant information that cannot be measured by the researchers, such as genetic information (biology), quality of life (economics), and wisdom (psychology). Disregarding that information could lead to wrong conclusions. For this reason, the main objective of this article is to propose a cure rate model by incorporating a structure of inflation or deflation of zeros among the initiated cells. In other words, $P(N = 0)$ is modeled by an extra parameter.

Hence, we assume that the number of competing causes of the event of interest follows an extension of the geometric distribution by including an additional parameter π (the ZMG distribution). The advantage of the proposed model is that the ZMG distribution is very flexible and has a natural interpretation in terms of both the "zero-inflated" or "zero-deflated" proportion of cure. Moreover, the proposed model also captures equidispersion, underdispersion, and overdispersion in the data, providing more flexibility in the cure rate model context. Furthermore, the cure proportion of the proposed model can be greater or lower than the cure proportion when the number of competing causes is modeled by the geometric distribution. Therefore, the ZMG distribution becomes the most appropriate alternative for modeling the number of competing causes of the event of interest in comparison with the usual discrete distributions already known in the literature.

This article is organized as follows: this is the introductory section. In Section 2, we introduce the proposed long-term survival model, discuss some of its properties, and present the estimation method for the model parameters. In Section 3, we conduct a Monte Carlo simulation study to evaluate the performance of the estimators. We also present and discuss two applications using real-world data sets. Finally, some conclusions are given in Section 4.

2 | ZMG CURE RATE REGRESSION MODEL

2.1 | Formulation

The ZMG distribution has been studied by some authors. Recently, Reference 35 introduced a first-order nonnegative integer-valued autoregressive process with zero-modified geometric innovations. We introduce the cure rate model where the number of competing causes is modeled by the ZMG distribution. In what follows, we shall briefly present the ZMG distribution. Let N be a discrete random variable (RV) following a ZMG distribution with parameters $\mu > 0$ and $\pi \in (-1/\mu, 1)$. More specifically, we assume that N has a probability mass function given by

$$\Pr(N = n; \mu, \pi) \equiv \Pr(N = n) = \begin{cases} \frac{1+\pi\mu}{1+\mu}, & \text{if } n = 0, \\ (1-\pi)\frac{\mu^n}{(1+\mu)^{n+1}}, & \text{if } n = 1, 2, \ldots \end{cases}. \tag{1}$$

We shall denote this distribution as ZMG(π, μ). The probability generating function (PGF) of N, denoted by $\varphi_N(s; \mu, \pi) = \mathrm{E}[s^N]$, is given by

$$\varphi_N(s; \mu, \pi) = \frac{1 + \pi\ \mu(1-s)}{1 + \mu(1-s)} = \pi + \frac{1-\pi}{1+\mu(1-s)}. \tag{2}$$

Equation (2) shows that, for $\pi \in [0, 1)$, the distribution of N is a mixture of a degenerate distribution at zero with mass π and a geometric distribution with mean μ and mass $1 - \pi$. For $\pi \in (0, 1)$ we have $N \overset{d}{=} N_1 \cdot N_2$, where N_1 and N_2 are independent random variables following a Bernoulli(π) distribution with mean $0 < \pi < 1$ and a geometric$(1/(1 + \mu))$ distribution with mean $\mu > 0$, respectively.

The mean and variance of N are

$$\mathrm{E}(N) \equiv \mu_N = \mu(1-\pi) \quad \text{and} \quad \mathrm{Var}(N) \equiv \sigma_N^2 = \mu(1-\pi)[1 + \mu(1+\pi)],$$

respectively. Thus, the dispersion index, which is the variance-to-mean ratio, is given by

$$I_N = \frac{\mu_N}{\sigma_N^2} = 1 + \mu(1+\pi).$$

Here, the model is appropriate for equidispersed data when $\pi = -1$, underdispersion when $\mu \in (0, 1)$ and $\pi \in [-1/\mu, -1)$, and overdispersion when $\pi \in (-1, 1)$. Furthermore, for $\pi \in (-1/\mu, 0)$ and $\pi \in (0, 1)$, we have a zero-deflated model and a zero-inflated model for the geometric distribution, respectively. Different values of π lead to different modifications of the ZMG distribution:

1. If $\pi = -1/\mu$, then the distribution in (1) becomes the zero-truncated geometric distribution, where the parameter π is not a model parameter, that is, there is no chance of getting a zero observation in the sample.
2. For $\pi \in (-1/\mu, 0)$, we have the zero-deflated geometric distribution, that is, a geometric distribution with a lower proportion of zeros.
3. If $\pi = 0$, then the corresponding distribution is the geometric distribution,
4. For $\pi \in (0, 1)$, we have the zero-inflated geometric distribution, which is a geometric process with a higher proportion of zeros.
5. If $\pi = 1$, then the corresponding zero-modified distribution is the one degenerated at zero.

Let N denote the number of competing causes related to the occurrence of an event of interest for an individual in the population (in a cancer context, N represents the carcinogenic cells of the individual). Assuming that $N \sim$ ZMG(π, μ) and conditional on $N = n$, we have that Y_j's are independent and identically distributed random variables representing the promotion times of the competing causes, for $j = 1, \ldots, n$, with survival function (SF) $S(\cdot; \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a vector of parameters. $S(\cdot; \boldsymbol{\xi})$ is proper in the sense that $\lim_{t\to\infty} S(t; \boldsymbol{\xi}) = 0$. Moreover, we assume that N is independent of $Y_1, \ldots, Y_n$, and the observable time-to-event is defined as $T = \min\{Y_1, \ldots, Y_N\}$ for $N \geq 1$, and $T = \infty$ for $N = 0$, leading to a cured fraction denoted by p_0.[10] For heterogeneous populations, we consider a set of p-dimensional vectors of covariates related

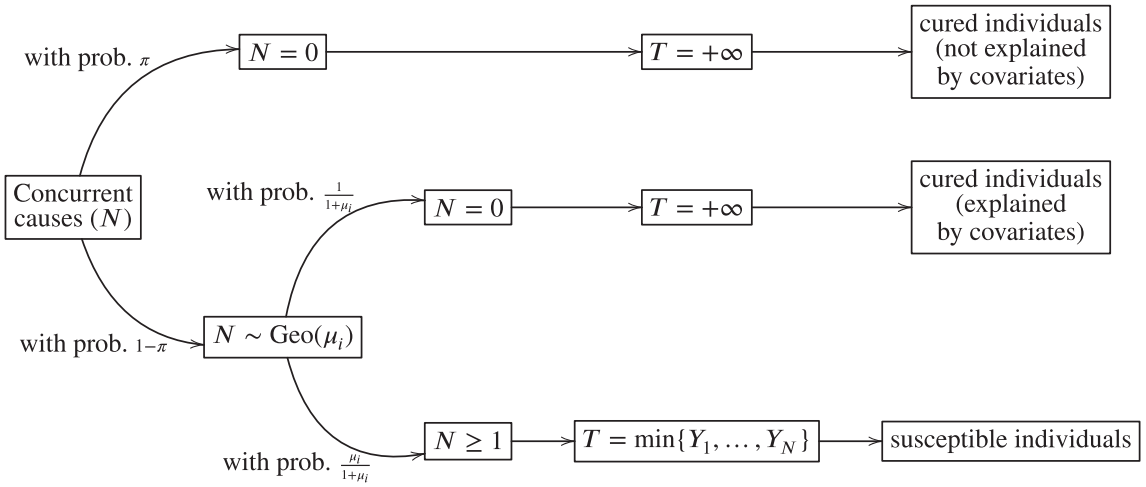

FIGURE 1 Relation for the time-to-event in cured and noncured individuals when $\pi \in [0, 1)$

to the cure rate, say $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ip})$, which can be introduced (conveniently) through the parameter μ as

$$\mu_i = \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right), \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of unknown parameters. Under this setting and for the cases where $\pi \in (0, 1)$, π is defined as the proportion of cured individuals in the population that cannot be explained by the covariates, that is, such individuals are cured by (unobservable) biological, social, economic, and so on, processes related to the problem or simply by nonmeasured information. On the other hand, in the $1 - \pi$ remaining of the population, the covariates can explain the underlying mechanism based on the geometric distribution for the number of concurrent causes, where they arise naturally susceptible individuals but also another part of cured individuals (see Equation (4)). Figure 1 illustrates this idea, where μ_i represents the mean of the individuals that can be explained by the covariates. Therefore, $\exp(\beta_k)$, $k = 1, \ldots, p$, can be interpreted as the relative increment (or decrement) to an observation that can be explained by covariates when its kth component is increased by one unit and the others are fixed.

Using (1), the probabilities of $N = 0$ and $N \geq 1$ are

$$\theta = \Pr(N = 0) = \frac{1 + \pi\ \mu}{1 + \mu} \quad \text{and} \quad 1 - \theta = \Pr(N \geq 1) = \frac{\mu(1 - \pi)}{1 + \mu},$$

respectively.

The run length of zeros of N, say M, follows a geometric distribution with termination probability θ, that is, $\Pr(M = m) = \theta^{m-1}(1 - \theta)$, $m \geq 1$. Thus, the average run length of zeros of M is given by $\mathrm{E}(M) = (1 + \mu)/\mu(1 - \pi)$. When the number of competing causes is modeled by the geometric distribution, M_0, the expected run length of zeros is $\mathrm{E}(M_0) = (1 + \mu)/\mu$. Note that $\mathrm{E}(M) = (1 - \pi)^{-1}\mathrm{E}(M_0)$. Thus, $E(M) \geq E(M_0)$ for $\pi \in [0, 1)$ and $E(M) < E(M_0)$ for $\pi \in (-1/\mu, 0)$.

Under this setup, the population SF, $S_{\text{pop}}(\cdot)$, can be computed as (see Reference 10 for details)

$$S_{\text{pop}}(t; \boldsymbol{\eta}) = \varphi_N(S(t; \boldsymbol{\xi}); \mu, \pi) = \frac{1 + \pi\ \mu\ F(t; \boldsymbol{\xi})}{1 + \mu\ F(t; \boldsymbol{\xi})}, \quad t > 0, \tag{3}$$

where $\boldsymbol{\eta} = (\mu, \pi, \boldsymbol{\xi})$ and $F(t; \boldsymbol{\xi}) = 1 - S(t; \boldsymbol{\xi})$. Thus,

$$S_{\text{pop}}(t; \boldsymbol{\eta}) = \pi + \frac{1 - \pi}{1 + \mu\ F(t; \boldsymbol{\xi})} = \pi + (1 - \pi) S^G_{\text{pop}}(t; \boldsymbol{\eta}), \quad t > 0, \tag{4}$$

where $S^G_{\text{pop}}(t; \boldsymbol{\eta}) = 1/[1 + \mu\ F(t; \boldsymbol{\xi})]$ corresponds to $S_{\text{pop}}(\cdot; \boldsymbol{\eta})$ when $\pi = 0$. Equation (4) shows that the population SF can be expressed as a mixture of a degenerate distribution at one (corresponding to the survival function of a "cured" individual) with mass π and the population SF, considering $N \sim \text{Geom}(1/(1 + \mu))$, with mass $1 - \pi$.

From (1) and (3), the cured fraction (p_0) is given by

$$p_0 = \lim_{t\to\infty} S_{\text{pop}}(t;\boldsymbol{\eta}) = \frac{1+\pi\ \mu}{1+\mu} = \pi + \frac{1-\pi}{1+\mu} = \pi + (1-\pi)p_0^G, \tag{5}$$

where p_0 denotes the proportion of "cured" or "immune" individuals in the population from which the sample was collected and $p_0^G = \lim_{t\to\infty} S_{\text{pop}}^G(t;\boldsymbol{\eta}) = 1/(1+\mu)$.

Consider p_0 and p_0^G the proportions of "cured" or "immune" individuals in the population when the number of competing causes is modeled by the ZMG and geometric distributions, respectively. Then, $p_0 \geq p_0^G$ if $\pi \in [0,1)$ and $p_0 < p_0^G$ if $\pi \in [1/\mu, 0)$. The above result shows that the proportion of cure, p_0, can be greater or lower than the proportion of cure when the number of competing causes is modeled by the geometric distribution.

The probability density function (PDF) associated to (3) reduces to

$$f_{\text{pop}}(t;\boldsymbol{\eta}) = \frac{\mu(1-\pi)f(t;\boldsymbol{\xi})}{[1+\mu\ F(t;\boldsymbol{\xi})]^2}, \quad t>0, \tag{6}$$

where $f(t;\boldsymbol{\xi}) = -dS(t;\boldsymbol{\xi})/dt$ denotes the (proper) PDF of the time-to-event T in (3) with hazard rate (HR) function

$$h_{\text{pop}}(t;\boldsymbol{\eta}) = \frac{\mu(1-\pi)f(t;\boldsymbol{\xi})}{[1+\mu\ F(t;\boldsymbol{\xi})][1+\pi\ \mu\ F(t;\boldsymbol{\xi})]} = \frac{(1-\pi)}{1+\pi\ \mu\ F(t;\boldsymbol{\xi})}\cdot h_{\text{pop}}^G(t;\boldsymbol{\eta}), \quad t>0, \tag{7}$$

where $h_{\text{pop}}^G(t;\boldsymbol{\eta}) = \mu\, f(t;\boldsymbol{\xi})/[1+\mu\ F(t;\boldsymbol{\xi})]$ corresponds to $h_{\text{pop}}(\cdot;\boldsymbol{\eta})$ when $\pi = 0$. Thus,

$$\lim_{t\to 0}\ h_{\text{pop}}(t;\boldsymbol{\eta}) = (1-\pi)\ \lim_{t\to 0}\ h_{\text{pop}}^G(t;\boldsymbol{\eta}) \quad \text{and} \quad \lim_{t\to\infty}\ h_{\text{pop}}(t;\boldsymbol{\eta}) = \frac{1-\pi}{1+\pi\ \mu}\ \lim_{t\to\infty}\ h_{\text{pop}}^G(t;\boldsymbol{\eta}).$$

It follows from (7) that

$$(1-\pi)\ h_{\text{pop}}^G(t;\boldsymbol{\eta}) \leq h_{\text{pop}}(t;\boldsymbol{\eta}) \leq \frac{1-\pi}{1+\pi\ \mu}\ h_{\text{pop}}^G(t;\boldsymbol{\eta}) \quad \text{for } \pi \in (-1/\mu, 0]$$

and

$$\frac{1-\pi}{1+\pi\ \mu}\ h_{pop}^G(t;\boldsymbol{\eta}) \leq h_{\text{pop}}(t;\boldsymbol{\eta}) \leq (1-\pi)\ h_{\text{pop}}^G(t;\boldsymbol{\eta}) \quad \text{for } \pi \in [0,1)\,.$$

The SF for the noncured population, say $S^*(t;\boldsymbol{\eta}) = \Pr(T>t|N\geq 1;\boldsymbol{\eta})$, is given by

$$S^*(t;\boldsymbol{\eta}) = \frac{S(t;\boldsymbol{\xi})}{[1+\mu\ F(t;\boldsymbol{\xi})]}, \quad t>0.$$

We note that $S^*(0;\boldsymbol{\eta}) = 1$ and $S^*(\infty;\boldsymbol{\eta}) = 0$, then we have that $S^*(t;\boldsymbol{\eta})$ is a proper SF. The PDF and HR functions related to the susceptible individuals reduce to

$$f^*(t;\boldsymbol{\eta}) = -\frac{dS^*(t)}{dt} = \frac{(1+\mu)f(t)}{[1+\mu\ F(t)]^2}, \quad t>0,$$

and

$$h^*(t;\boldsymbol{\eta}) = \frac{(1+\mu)f(t;\boldsymbol{\xi})}{S(t;\boldsymbol{\xi})[1+\mu\ F(t;\boldsymbol{\xi})]^2} = \frac{1+\mu}{[1+\mu\ F(t;\boldsymbol{\xi})]^2}\cdot h(t;\boldsymbol{\xi}), \quad t>0.$$

Note that the parameters π and μ control the cure rate of the model (see Equation (5)). However, the distribution related to the noncured individuals does not depend on π.

2.2 | Identifiability of the model

Identifiability of cure rate models is a critical point. Reference 19 discussed the identifiability of the mixture model in Reference 2 and the promotion time cure rate model in References 8,9, and Reference 20 revisited such discussion in a

more general framework for the cure rate models. The ZMG cure rate model is identifiable if $\boldsymbol{\eta}_1 \neq \boldsymbol{\eta}_2$ implies $S_{\text{pop}}(t;\boldsymbol{\eta}_1) \neq S_{\text{pop}}(t;\boldsymbol{\eta}_2)$, $\forall t > 0$. In particular, taking $t = +\infty$ implies that the cure rate generated by $\boldsymbol{\xi}_1 \neq \boldsymbol{\xi}_2$ is different. As in any cure rate model, if the cure rate is not modeled by covariates then the model is not identifiable. This can be seen taking $\boldsymbol{\xi}_1 = \left(\pi = \frac{q(1+\mu)-1}{\mu}, \mu\right)$ and $\boldsymbol{\xi}_2 = \left(\pi, \mu = \frac{1-q}{q-\pi}\right)$, for any $q \in (0,1)$, $\mu > 0$, and $\pi \in (-1/\mu, 1)$, because both pairs generate the same $p_0 = q$.

Suppose that a (nonnull) set of covariates $\boldsymbol{x}$ are available, which are considered through $\mu = \exp(\boldsymbol{x}^\top \boldsymbol{\beta})$. The population survival function in (3) can also be written as

$$S_{\text{pop}}(t;\boldsymbol{\eta}) = \underbrace{\frac{1+\pi\mu}{1+\mu}}_{p_0(\boldsymbol{\beta},\pi)} + \underbrace{\frac{\mu(1-\pi)}{(1+\mu)}}_{1-p_0(\boldsymbol{\beta},\pi)} \times \underbrace{\frac{(1+\mu)}{(1+\mu F(t;\boldsymbol{\xi}))}}_{S^*(t;\boldsymbol{\beta},\boldsymbol{\xi})}.$$

Therefore, the ZMG cure rate model can be seen as a mixture model with cure rate $p_0(\boldsymbol{\beta},\pi)$ and survival function for the time-to-event in the susceptible individuals $S^*(t;\boldsymbol{\beta},\boldsymbol{\xi})$. Let $||p_0||_\beta = \sup\{p_0(\boldsymbol{\beta},\pi) : \pi \in (-1/\mu, 1]\}$. It can be shown that $||p_0||_\beta = 1$. Therefore, by point 1 from theorem 1 in Reference 20, the ZMG cure rate model (considering covariates through the parameter μ) is identifiable, independently from the specified form to $F(\cdot;\boldsymbol{\xi})$.

There is a vast literature regarding parametric approaches in cure rate models. To name a few, References 2-4,36 for the mixture model, References 9,37 for the promotion time cure rate model, and References 13,18,38 for other cure rate models. However, advances in nonparametric techniques are scarce in the literature. This is explained (partially) by the difficulty to propose a nonparametric estimator in each cure rate model. For instance, in the mixture model context, References 39-41 considered a nonparametric estimator for the failure time distribution of uncured individuals based on proportional hazard models and the Kaplan-Meier (KM) estimator, respectively. In the promotion time cure rate model, Reference 42 proposed to model F (which in this case represents the failure time distribution of the concurrent causes) using P-splines.

On the other hand, for many biological problems, the Weibull model is enough to be considered for this function (we shall refer to this as the ZMGWEcr model). As we will show in the application section, the ZMGWEcr model is a very competitive model in many practical situations. However, we leave the option open to use the ZMG under other distributions with positive support or even a nonparametric structure.

2.3 | Estimation

In order to estimate the model parameters, we utilize the maximum likelihood (ML) method. We consider the situation when the time to an event is not completely observed and is subject to right censoring. Let c_i be the censoring time for the ith individual. We observe $t_i = \min(y_i, c_i)$ and $\delta_i = \mathbb{I}(y_i \leq c_i)$, where $\delta_i = 1$ if y_i is a time-to-event and $\delta_i = 0$ if y_i is right-censored, $i = 1, \ldots, n$. Based on n observed vectors $(y_1, \delta_1, \boldsymbol{x}_1^\top), \ldots, (y_1, \delta_n, \boldsymbol{x}_n^\top)$, the corresponding likelihood function under uninformative censoring is expressed as

$$L(\boldsymbol{y};\boldsymbol{\vartheta}) = \prod_{i=1}^{n} [f_{\text{pop}}(y_i;\boldsymbol{\vartheta})]^{\delta_i} [S_{\text{pop}}(y_i;\boldsymbol{\vartheta})]^{1-\delta_i}, \tag{8}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\xi}, \boldsymbol{\beta})^\top$, $S_{\text{pop}}(y_i;\boldsymbol{\vartheta})$, and $f_{\text{pop}}(y_i;\boldsymbol{\vartheta})$ are given in (3) and (6), respectively. To model the time-to-event for the concurrent causes, we can consider any asymmetric distribution. In the simulation study, we consider the Weibull (WE) distribution for illustrative purposes. Henceforth, using the SF in Equation (3) together with the WE distribution we have ZMGWEcr model. The PDF and CDF functions for the WE distribution, denoted by $T \sim \text{WE}(\lambda, \gamma)$, are given by $f(t;\lambda,\gamma) = \lambda\ \gamma\ t^{\gamma-1} \text{e}^{-\lambda\ t^\gamma}$ and $S(t;\lambda,\gamma) = \exp(-\lambda\ t^\gamma)$, respectively, where $\gamma > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution.

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. We can easily perform the maximum likelihood estimation by using standard nonlinear maximization procedures found in most statistical and data analysis packages. Here, we use the `R` software by its functions `optim` (or `optimx`); see www.r-project.org and Reference 43. Such procedures require initial guesses for the parameters to be initialized. As $\pi = 0$ corresponds to the geometric cure rate model (see

References 36 for details), a good alternative to initialize the procedure for the ZMGWEcr model (or any distribution for the time-to-event of the concurrent causes) is to take the ML estimators of the geometric cure rate model with the same distribution for $S(\cdot;\boldsymbol{\xi})$ (WE or any other). A detailed procedure for obtaining the ML estimators in such model is described in Reference 44.

3 | NUMERICAL APPLICATIONS

In this section, we shall report the results of a Monte Carlo (MC) simulation study we conducted to evaluate the performance of the ML estimators of the proposed model parameters. We assume that the time-to-event of interest follows a WE distribution. We also show the applicability of the proposal by presenting and discussing two applications using real-world medical data sets.

3.1 | A simulation study

For the time-to-event for the concurrent causes, we consider the WE model with parameters $\lambda = 1$ and $\gamma = 2$. The simulation scenario considers sample sizes $n \in \{200, 400, 1000, 2000\}$, values of parameter $\pi \in \{-0.20, 0.20\}$ and 5000 MC replications. For each individual, we also consider a covariate x generated from a Bernoulli distribution with parameter 0.5, which was included in $\mu_i = \exp(\beta_0 + \beta_1 x_i)$. Also, $\beta_0 = 0.50$ and $\beta_1 = -1.00$, producing a cure fraction for the two levels of x, say p_{00} and p_{01} for the levels $x = 0$ and $x = 1$, respectively. Such values are $p_{00} = 0.253$ and $p_{01} = 0.547$ for $\pi = -0.2$, and $p_{00} = 0.502$ and $p_{01} = 0.698$ for $\pi = 0.2$. We used Algorithm 1 to generate observed times and censoring indicators.

Algorithm 1. Generator of random numbers from the ZMG-G model

1: Fix the values of the parameters and of the cure fraction for each group p_0 and p_1;
2: Obtain a random number of the covariate $\boldsymbol{x}$ from $X \sim$ Bernoulli(1/2); if the covariate is 0,then generate $W_i \sim$ Bernoulli$(1 - p_0)$;
3: If $W_i = 0, y^* = \infty$. If $W_i = 1$, then y^* is the root of $G(y) = u$, where $u \sim$ Uniform$(0, 1 - p_0)$. Here, $G(\cdot)$ is the cumulative distribution function of the baseline distribution chosen for the time of occurrence of an event of interest;
4: Repeat steps 1 to 3 using p_1 instead of p_0. If in the third step the value of the covariate is 1 go to step 5;
5: Generate $u_i^* \sim$ Uniform$(0, \max(y_i^*))$, considering only finite values of y_i;
6: Obtain $y_i = \min(y_i, u_i)$. If $y_i > u_i$, then $\delta_i = 1$, otherwise $\delta_i = 0$;
7: Repeat steps 1 to 6 until the required number of data has been generated.

Table 1 shows the empirical mean (EM) and the mean squared error (MSE) of the ML estimators for each value of the parameters and sample sizes considered. Note that, as the sample size increases, the ML estimators of the cured fraction become more efficient. In general, all of these results show the good performance of the proposed model.

3.2 | Applications to medical modeling

In this subsection, we shall apply the proposed model in two real data sets. Because of the genesis of the Weibull distribution, the medical data are by excellence ideally modeled by this distribution.[13,17] Thus, the use of the Weibull distribution for fitting these two data sets is well justified. The required numerical evaluations for the data analysis were performed using the `R` software.

3.2.1 | Melanoma data

The first real-world medical data set corresponds to the survival time after an operation for removal of a malignant melanoma (ranging from 0.0274 to 15.25 years) until the patient's death, which is possibly censored. The data set contains $n = 205$ patients and is available in the `timereg` package of the `R` software; www.r-project.org. The explanatory

TABLE 1 MC simulation results of the ML estimators (MSE in parentheses) for the ZMGWEcr model

Sample sizes	EM						
(censoring rates)	$\hat{\pi}$	$\hat{\lambda}$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{p}_{00}$	$\hat{p}_{01}$
$(\pi, \lambda, \gamma, \beta_0, \beta_1, p_{00}, p_{01}) = (-0.2, 1, 2, 0.5, -1.0, 0.2530, 0.5470)$							
200 (49.05%)	−0.5852	1.0302	2.5674	0.5528	−1.0626	0.2516	0.5354
	(1.9553)	(0.1152)	(2.1930)	(1.3160)	(0.3497)	(0.0564)	(0.0730)
400 (47.77%)	−0.4927	1.0058	2.0570	0.3782	−0.9944	0.2533	0.5447
	(0.9422)	(0.0781)	(0.5504)	(0.7966)	(0.2428)	(0.0366)	(0.0428)
1000 (46.64%)	−0.2663	1.0021	2.0077	0.4534	−0.9919	0.2539	0.5456
	(0.2571)	(0.0481)	(0.2929)	(0.4265)	(0.1532)	(0.0222)	(0.0254)
2000 (45.92%)	−0.2205	1.0027	2.0123	0.4972	−1.0035	0.2530	0.5461
	(0.1439)	(0.0334)	(0.1956)	(0.2802)	(0.1079)	(0.0154)	(0.0175)
$(\pi, \lambda, \gamma, \beta_0, \beta_1, p_{00}, p_{01}) = (0.2, 1, 2, 0.5, -1.0, 0.502, 0.698)$							
200 (66.62%)	0.5558	1.0278	3.1179	0.4084	−1.0556	0.4986	0.6892
	(3.7318)	(0.144)	(5.0235)	(2.0062)	(0.4370)	(0.0623)	(0.0708)
400 (65.82%)	0.1072	1.0162	2.2926	0.4185	−1.0134	0.5030	0.6938
	(1.4043)	(0.0999)	(1.3125)	(1.1845)	(0.3123)	(0.0408)	(0.0419)
1000 (64.65%)	0.1135	1.0068	2.0646	0.4655	−1.0003	0.5026	0.6956
	(0.3162)	(0.0601)	(0.4002)	(0.5630)	(0.1946)	(0.0245)	(0.0236)
2000 (64.18%)	0.1738	1.0023	2.0147	0.4850	−1.0012	0.5022	0.6973
	(0.1356)	(0.0416)	(0.2518)	(0.3597)	0.1373	(0.0171)	(0.0161)

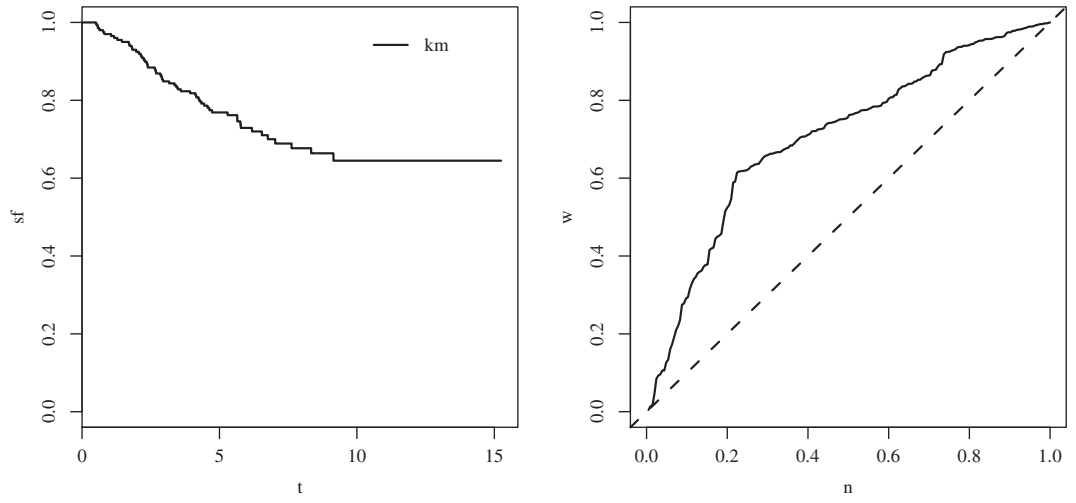

FIGURE 2 Fitted SF (left) and TTT plot (right) for melanoma data

variables are: tumor thickness (x_{i1}, in mm, mean = 2.92 and SD = 2.96); ulceration (x_{i2}, absent [115 patients] and present [90 patients]); and gender (x_{i3}, male [79 patients] and female [126 patients]). Figure 2 (left) shows the KM estimate of the SF for the malignant melanoma data. We observe that the estimated curve stabilizes in the levels above 0.65, suggesting that the patients censored at the end of the experiment may be immune to the risk in question or were cured during the experiment.

A descriptive summary of the observed times (in years) provides the following sample values: median = 5.49, mean = 5.90, SD = 3.07, coefficient of variation (CV) = 52.12, coefficient of skewness (CS) = 0.33, coefficient of

TABLE 2 ML estimates (corresponding estimated asymptotic SE in parentheses) for the MWEcr, GWEcr, and ZMGWEcr models along with model selection measures and respective *P*-values in brackets; melanoma data

Estimate	MWEcr	GWEcr	ZMGWEcr
$\hat{\pi}$	—	—	0.324 (0.121)
$\hat{\lambda}$	1.599 (0.204)	1.938 (0.235)	2.189 (0.309)
$\hat{\gamma}$	4.726 (0.631)	6.219 (1.169)	7.487 (2.039)
$\hat{\beta}_0$	1.866 (0.338)	−1.915 (0.350)	−1.223 (0.595)
$\hat{\beta}_1$	−0.166 (0.078)	0.178 (0.054)	0.240 (0.073)
P-value	[.034]	[.001]	[0.001]
$\hat{\beta}_2$	−1.498 (0.416)	1.596 (0.356)	1.871 (0.483)
P-value	[<.001]	[<.001]	[<.001]
AIC	431.076	421.003	**419.346**
BIC	447.671	437.618	439.283

kurtosis (CK) = −0.28, whereas their minimum and maximum times are 0.0274 and 15.25, respectively. From this summary, we observe the positively skewed nature and the moderate kurtosis level of the data distribution. The identification of the shape of an HR can be obtained by the scaled total time on test (TTT) function.[45] Figure 2 (right) suggests an increasing HR for the observed times. Therefore, the Weibull distribution is a good candidate since it allows the modeling of constant, increasing, and decreasing HRs. Also, the Weibull distribution is one of the most used distributions in survival and reliability analysis because of its good properties and flexibility in data modeling.

Table 2 provides the ML estimates with standard errors (SE) for the mixture Weibull cure rate (MWEcr), geometric Weibull cure rate (GWEcr), and ZMGWEcr models, along with two goodness-of-fit statistics: Akaike information criterion (AIC) and Bayesian information criterion (BIC).[46,47] It is noteworthy that the ZMGWEcr model nests the other models. Notice that the ZMGWEcr model presents the smallest AIC values, suggesting that this model provides the best fit to these data. We also highlight that 32.4% of the population is cured after the removal of the malignant melanoma by causes that cannot be explained by the presence/absence of ulceration and tumor thickness. Such percentage is approximately half of the cured individuals visualized in the KM estimator in Figure 2, where covariates are ignored. Differently, for individuals where covariates can explain the underlying latent process, the expected value for the number of carcinogenic cells increases $\exp(0.239) \approx 1.27$ times for each cm that increases the tumor and $\exp(1.868) \approx 6.48$ times for individuals with ulceration, when compared with individuals without ulceration. We test the null hypothesis $\mathcal{H}_0$: GWEcr against the alternative hypothesis $\mathcal{H}_1$: ZMGWEcr, that i s, $\mathcal{H}_0 : \pi = 0$ against $\mathcal{H}_1 : \pi \neq 0$. We obtained a likelihood ratio (LR) test statistic equal to 3.657 (*P*-value = .056). Thus, the null hypothesis is rejected at 10% of significance level, favoring the ZMGWEcr model. Figure 3 contains quantile-quantile (QQ) plot with normalized randomized quantile (RQ) residuals[48] for the ZMGWEcr model, indicating that the normalized randomized quantile residuals present a good agreement with the $\mathcal{N}(0, 1)$ distribution.

3.2.2 | Blood and marrow transplantation data

The second real-world data set comes from the European Society for Blood and Marrow Transplantation (EBMT) and corresponds to the time in years from transplantation to relapse (ranging from 0.00082 to 17.26 years). The data set contains 2279 patients transplanted at the EBMT between 1985 and 1998. The data set is available in the `mstate` package of the `R` software. We considered the covariate that indicates the presence/absence of Prophylaxis (x_{i1}, yes [540 patients] and no [1730 patients]); descriptive statistics of the observed lifetimes from the mentioned data are median = 3.46, mean = 4.48, SD = 4.21, CV = 93.82, CK = −0.46, and CS = 0.76, indicating the positively skewed nature and high kurtosis level of the data distribution, whereas their minimum and maximum times are 0.0082 and 17.26, respectively. Figure 4 shows the fitted SF by KM and the TTT plot for these data. Note that, according to Figure 4 (left), the estimated curve stabilizes in the levels above 0.75, indicating that the censored patients at the end of the experiment may be immune to the risk in question or were cured during the experiment. The TTT plot displayed in Figure 4

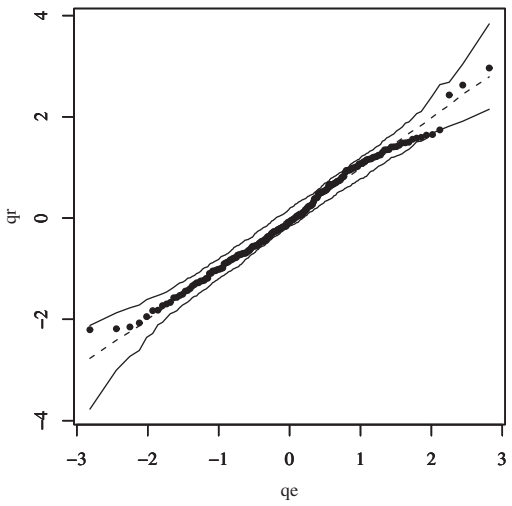

FIGURE 3 Quantile-quantile (QQ) plot with envelope for RQ residuals for the ZMGWEcr model; melanoma data

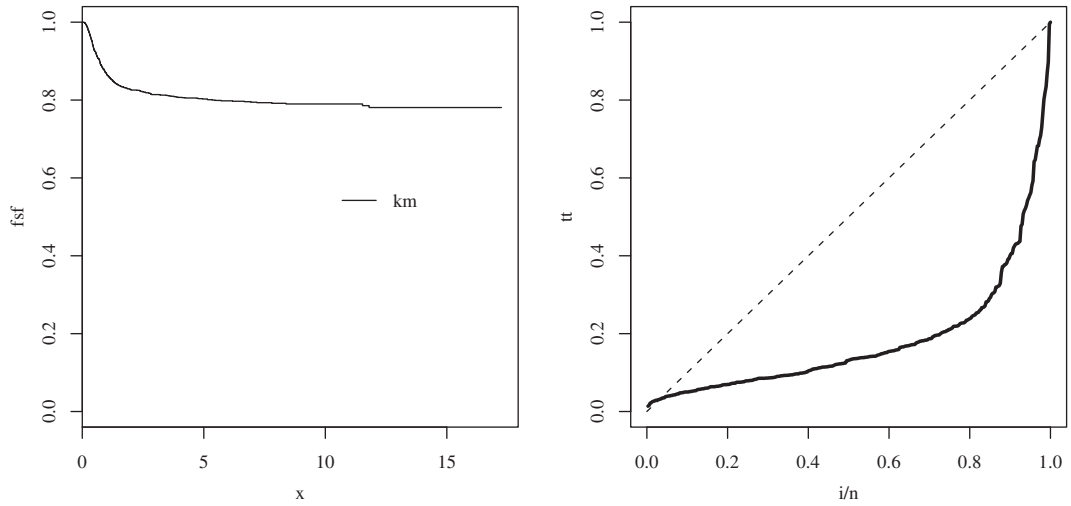

FIGURE 4 Fitted SF (left) and TTT plot (right) for EBMT4 data

(right) suggests an increasing HR for the observed lifetimes, which again justifies the use of the Weibull distribution as a baseline HR.

The ML estimates with corresponding SE of the model parameters and model selection measures are reported in Table 3. The ZMGWEcr regression model presents the smallest AIC and BIC values, suggesting that the proposed model provides the best fit to this data set. In addition, from the LR test, we obtain an LR test statistic equal to 82.26 (P-value $< .001$), favoring the ZMGWEcr model. Figure 5 displays the QQ plot with normalized randomized quantile residuals for the ZMGWEcr model. This graphical plot shows the notorious agreement, in terms of fitting to the data, of the ZMGWEcr regression model. On the other hand, 77.9% of the cured individuals (almost all of the cured individuals according to the KM estimator in Figure 4) is not explained by the prophylaxis. Finally, for patients where the prophylaxis can explain the subjacent process, the mean of concurrent causes for individuals with prophylaxis is $\exp(-0.506) \approx 0.60$ times the mean of concurrent causes for individuals without prophylaxis.

4 | CONCLUDING REMARKS

In this article, we proposed a new model for survival data assuming competing causes of the event of interest and following the zero-modified geometric distribution. More specifically, we assumed that N belongs to a zero-modified distribution, which includes the geometric distribution as a particular case. The ZMG distribution is an appropriate

FIGURE 5 Quantile-quantile (QQ) plot with envelope for RQ residuals for the ZMGWEcr model; blood and marrow transplantation data

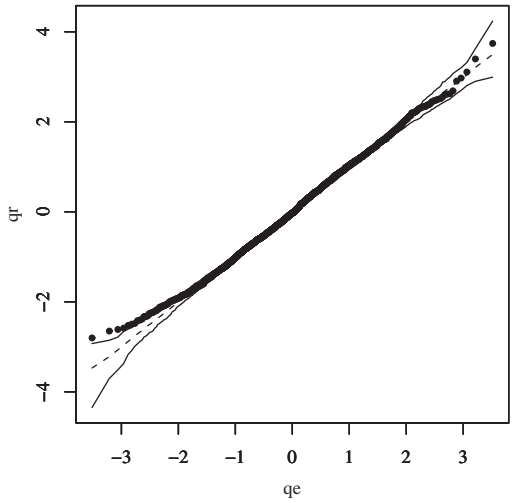

TABLE 3 ML estimates (corresponding estimated asymptotic SE in parentheses) for the MWEcr, GWEcr, and ZMGWEcr models along with model selection measures and respective *P*-values in brackets; blood and marrow transplantation data

Estimate	MWEcr	GWEcr	ZMGWEcr
$\hat{\pi}$	—	—	0.786 (0.014)
$\hat{\lambda}$	1.027 (0.038)	1.085 (0.040)	1.707 (0.084)
$\hat{\gamma}$	1.332 (0.077)	1.547 (0.092)	5.615 (0.322)
$\hat{\beta}_0$	1.118 (0.115)	−1.084 (0.113)	3.099 (0.338)
$\hat{\beta}_1$	0.287 (0.134)	−0.308 (0.129)	0.687 (0.200)
P-value	[.032]	[.017]	[<.001]
AIC	2671.464	2662.165	**2581.907**
BIC	2694.390	2685.091	**2610.564**

alternative for modeling clustered samples when the number of competing causes of the event of interest consists of two subpopulations, one containing only zeros (cure proportion), while in the other (noncure proportion) the number of competing causes of the event of interest follows a geometric distribution. The advantage of this assumption is that we can measure the proportion of cure in the initiated cells. We discussed estimation and associated inference by maximum likelihood. In the empirical applications, we discovered that the new cure rate model provided the best fit among other common models in the literature. We hope the proposed model attracts the attention of practitioners in survival analysis.

ORCID

Jeremias Leão https://orcid.org/0000-0003-1176-0198
Marcelo Bourguignon https://orcid.org/0000-0002-1182-5193
Diego I. Gallardo https://orcid.org/0000-0001-8184-7403
Ricardo Rocha https://orcid.org/0000-0003-3057-9482
Vera Tomazella https://orcid.org/0000-0002-6780-2089

REFERENCES

1. Cox DR. Regression models and life tables. *J Royal Stat Soc Ser B*. 1972;34(2):187-220.
2. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *J Am Stat Assoc*. 1952;47(259):501-515.
3. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*. 1982;38(4):1041-1046.
4. Maller RA, Zhou X. *Survival Analysis with Long-Term Survivors*. New York, NY: Wiley; 1996.
5. Ewell M, Ibrahim JG. The large sample distribution of the weighted log rank statistic under general local alternatives. *Lifetime Data Anal*. 1997;3(1):5-12.
6. Stangl DK, Greenhouse JB. Assessing placebo response using Bayesian hierarchical survival models. *Lifetime Data Anal*. 1998;4(1):5-28.
7. Calsavara VF, Tomazella VLD, Fogo JC. The effect of frailty term in the standard mixture model. *Chilean J Stat*. 2013;4(2):95-109.

8. Yakovlev AY, Tsodikov AD. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Hackensack, NJ: World Scientific; 1996.
9. Chen M-H, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. *J Am Stat Assoc*. 1999;94(447):909-919.
10. Rodrigues J, Cancho VG, Castro M, Louzada-Neto F. On the unification of long-term survival models. *Stat Probab Lett*. 2009;79(6):753-759.
11. Rodrigues J, De Castro M, Cancho VG, Balakrishnan N. COM–Poisson cure rate survival models and an application to a cutaneous melanoma data. *J Stat Plann Infer*. 2009;139(10):3605-3611.
12. Cancho VG, Louzada F, Ortega EM. The power series cure rate model: an application to a cutaneous melanoma data. *Commun Stat Simulat Comput*. 2013;42(3):586-602.
13. Gallardo DI, Gómez HW, Bolfarine H. A new cure rate model based on the Yule–Simon distribution with application to a melanoma data set. *J Appl Stat*. 2017;44(7):1153-1164.
14. Cooner F, Banerjee S, Carlin BP, Sinha D. Flexible cure rate modeling under latent activation schemes. *J Am Stat Assoc*. 2007;102(478):560-572.
15. Borges P, Rodrigues J, Louzada F, Balakrishnan N. A cure rate survival model under a hybrid latent activation scheme. *Stat Methods Med Res*. 2016;25(2):838-856.
16. Cancho VG, Castro M, Dey DK. Long-term survival models with latent activation under a flexible family of distributions. *Brazilian J Probab Stat*. 2013;27(4):585-600.
17. Rodrigues J, Castro M, Balakrishnan N, Cancho VG. Destructive weighted Poisson cure rate models. *Lifetime Data Anal*. 2011;17(3):333-346.
18. Cancho VG, Bandyopadhyay D, Louzada F, Yiqi B. The destructive negative binomial cure rate model with a latent activation scheme. *Stat Methodol*. 2013;13:48-68.
19. Li C-S, Taylor JMG, Sy-Judy P. Identifiability of cure models. *Stat Probab Lett*. 2001;54(4):389-395.
20. Hanin L, Huang LS. Identifiability of cure models revisited. *J Multivar Anal*. 2014;130(1):261-274.
21. Caroni C, Crowder MJ, Kimber AC. Discrete frailty models for heterogeneity in lifetime distributions. In: Stasinopoulos M, Touloumi G, eds. *Proceedings of the 17th International Workshop on Statistical Modelling*; 2002a:157-162.
22. Caroni C, Crowder MJ, Kimber AC. Discrete frailty. In: Langseth H, Lindqvist B, eds. *Proceedings of the 3rd International Conference on Mathematical Methods in Reliability MMR 2002*. Trondheim, Norway: NTNU; 2002b:135-138.
23. Caroni C, Crowder M, Kimber A. Proportional hazards models with discrete frailty. *Lifetime Data Anal*. 2010;16(3):374-384.
24. Ata N, Özel G. Survival functions for the frailty models based on the discrete compound Poisson process. *J Stat Comput Simul*. 2013;83(11):2105-2116.
25. Yin G. Bayesian cure rate frailty models with application to a root canal therapy study. *Biometrics*. 2005;61(2):552-558.
26. Souza D, Cancho VG, Rodrigues J, Balakrishnan N. Bayesian cure rate models induced by frailty in survival analysis. *Stat Methods Med Res*. 2017;26(5):2011-2028.
27. Leão J, Leiva Víctor, Saulo H, Tomazella V. Incorporation of frailties into a cure rate regression model and its diagnostics and application to melanoma data. *Stat Med*. 2018;37(29):4421-4440.
28. Cancho VG, Barriga G, Leão J, Saulo H. Survival model induced by discrete frailty for modeling of lifetime data with long-term survivors and change-point. *Commun Stat Theory Methods*. 2019;1-12. https://www.tandfonline.com/doi/abs/10.1080/03610926.2019.1648826.
29. Balka J, Desmond AF, McNicholas PD. Review and implementation of cure models based on first hitting times for Wiener processes. *Lifetime Data Anal*. 2009;15(2):147.
30. Balka J, Desmond AF, McNicholas PD. Bayesian and likelihood inference for cure rates based on defective inverse Gaussian regression models. *J Appl Stat*. 2011;38(1):127-144.
31. Rocha R, Nadarajah S, Tomazella V, Louzada F. Two new defective distributions based on the Marshall–Olkin extension. *Lifetime Data Anal*. 2016;22(2):216-240.
32. Rocha R, Nadarajah S, Tomazella V, Louzada F. A new class of defective models based on the Marshall–Olkin family of distributions for cure rate modeling. *Comput Stat Data Anal*. 2017;107:48-63.
33. Rocha R, Nadarajah S, Tomazella V, Louzada F, Eudes A. New defective models based on the Kumaraswamy family of distributions with application to cancer data sets. *Stat Methods Med Res*. 2017;26(4):1737-1755.
34. Scudilio J, Calsavara VF, Rocha R, Louzada F, Tomazella V, Rodrigues AS. Defective models induced by gamma frailty term for survival data with cured fraction. *J Appl Stat*. 2019;46(3):484-507.
35. Bourguignon M. Modelling time series of counts with deflation or inflation of zeros. *Stat Interf*. 2018;11(4):631-639.
36. Cancho VG, Louzada F, Barriga GDC. The geometric Birnbaum–Saunders regression model with cure rate. *J Stat Plann Infer*. 2012;142(4):993-1000.
37. Ortega EMM, Cancho VG, Paula A. Generalized log-gamma regression models with cure fraction. *Lifetime Data Anal*. 2009;15(1):79-106.
38. De Castro M, Cancho VG, Rodrigues J. A Bayesian long-term survival model parametrized in the cured fraction. *Biometr J J Math Methods Biosci*. 2009;51(3):443-455.
39. Kuk AYC, Chen C-H. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*. 1992;79(3):531-541.
40. Peng Y, Dear KBG. A nonparametric mixture model for cure rate estimation. *Biometrics*. 2008;56(1):237-243.
41. Taylor JMG. Semi-parametric estimation in failure time mixture models. *Biometrics*. 1995;51(3):899-907.

42. Lambert P, Bremhorst V. Estimation and identification issues in the promotion time cure model when the same covariates influence long- and short-term survival. *Biom J*. 2019;61(2):275-289.
43. R-Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.
44. Gallardo DI, Romeo JS, Meyer R. A simplified estimation procedure based on the EM algorithm for the power series cure rate model. *Commun Stat Simulat Comput*. 2017;46(8):6342-6359.
45. Aarset MV. How to identify a bathtub hazard rate. *IEEE Trans Reliab*. 1987;36(1):106-108.
46. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, eds. *Proceedings of the 2nd International Symposium on Information Theory*. Budapest: Akademiai Kiado; 1973:267-281.
47. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6:461-464.
48. Dunn PK, Smyth GK. Randomized quantile residuals. *J Comput Graph Stat*. 1996;5:236-244.

How to cite this article: Leão J, Bourguignon M, Gallardo DI, Rocha R, Tomazella V. A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. *Statistics in Medicine*. 2020;1–13. https://doi.org/10.1002/sim.8664