Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

A defective cure rate quantile regression model for male breast cancer data[†]

Agatha Rodrigues*[1,2] | Patrick Borges[1] | Bruno Santos[3]

[1]Department of Statistics, Federal University of Espírito Santo, Espírito Santo, Brazil
[2]Division of Clinical Obstetrics, Hospital das Clinicas HCFMUSP, Faculty of Medicine, University of São Paulo, São Paulo, Brazil
[3]School of Mathematics, Statistics and Actuarial Science, University of Kent, Kent, United Kingdom

Correspondence
*Agatha Rodrigues, Email: agatha.rodrigues@ufes.br

Summary

In this article, we particularly address the problem of assessing the impact of different prognostic factors, such as clinical stage and age, on the specific survival times of men with breast cancer when cure is a possibility, where there is also the interest of explaining this impact on different quantiles of the survival times. To this end, we developed a quantile regression model for survival data in the presence of long-term survivors based on the generalized distribution of Gompertz in a defective version, which is conveniently reparametrized in terms of the q-th quantile and then linked to covariates via a logarithm link function. This proposal allows us to obtain how each variable affects the survival times in different quantiles. In addition, we are able to study the effects of covariates on the cure rate as well. We consider Markov Chain Monte Carlo methods to develop a Bayesian analysis in the proposed model and we evaluate its performance through a Monte Carlo simulation study. Finally, we illustrate the advantages of our model in a data set about male breast cancer from Brazil.

KEYWORDS:
Cure fraction, Defective distribution, Generalized Gompertz distribution, Quantile regression

1 | INTRODUCTION

1.1 | Bibliographical review

In population-based cancer survival studies, survival models that take into account a cure fraction, known as cure rate models or long-term survival models, are constantly used to simultaneously explain immune and susceptible patients to the cause of failure event of interest under study. The most common approach to the cure rate model is the standard mixture model approached by[2] and extended by[3] and later extensively studied by[4,5,6,7,8,9], among others. Although the mixture cure rate model is the most widely used in the literature, other alternative models of population modeling with cured individuals have been studied, such as: (i) models based on the latent competitive risk structure[10]; (ii) models based on defective distributions, defined as distributions that are not normalized to one for some values of its parameters, a concept discussed by[11].

Based on item (i), for example,[12] and[13] in cancer recurrence scenarios, assume that a latent biological process of propagation of latent clonogenic tumor cells (latent competitive risks) produces the observed failure (relapse), resulting in the model known in the literature as bounded cumulative hazard model. A unified approach of the standard mixture and bounded cumulative hazard models was proposed by[14].[15] developed a more flexible cure rate survival model which includes a destructive process of the initial risk factors in a competitive scenario and is thus based on the biological mechanism of the occurrence of the event of

[†]A defective cure rate quantile regression model for male breast cancer data.

interest. Subsequently,[16] proposed a new Bayesian flexible cure rate survival model which generalizes the stochastic model of[17] and has much in common with the destructive cure rate model formulated by[15]. [18] introduced a new cure rate survival model which extends the model of[15] by incorporating a structure of dependence between the initiated cells. Recently,[19] proposed a cure rate survival hybrid model for accommodating characteristics of unobservable stages (initiation, promotion and progression) of carcinogenesis from survival data in the presence of latent competing risks.

Regarding the item (ii), defective distributions have the advantage of allowing a cure rate without adding any extra parameters in the modeling, unlike the above-mentioned cure rate models[20]. They are obtained from standard distributions by changing the domain of the parameters in such a way that $\lim_{t\to\infty} S(t) = p_0 \in (0, 1)$. The two main distributions used for this purpose are the inverse Gaussian and Gompertz. Proposals dealing with the defective distributions have appeared in the literature, to name a few we refer to[21,22,23,24,25,26].

For all of the aforementioned survival models, inferential results are evaluated in terms of hazard ratio or on the average of the logarithm of the failure times, not directly on the failure time. However, physicians can be interested in assessing earlier stage of the follow-up, for example, and quantile regression can directly model the lower quantile of interest to provide the covariate effects specifically for this quantile. Under this approach, it is possible to know how the regression coefficients of a given covariate change for different quantiles of survival time.

Quantile regression, introduced by[30], is particularly useful when the rate of change in the conditional quantiles, expressed by the regression coefficients, depends on the quantile. One of its main advantages concerns its flexibility for modeling data with heteroscedasticity. This feature makes quantile regression very attractive, because the failure times often are quite right-skewed[34].

Even though quantile regression for survival data has been extensively studied in the literature (e.g.[31,32,33,34,35,37,38,35]), there are not many references dealing with quantile regression in a cure rate setup, known as cure rate quantile regression models;[39] and[29] could be mentioned as some of the examples. These proposals have included a non-parametric component either in the functional form of the regression equation or the distribution of the lifetime of the model, or both.

However, to the best of our knowledge, no one has so far examined a particular fully parametric approach to cure rate quantile regression based on defective distributions, which, as mentioned earlier, have the competitive advantage of allowing a cure rate without adding any extra parameters in the modeling process, in contrast to the standard mixture approach considered by[39]. Therefore, the main objective of this paper is to consider a quantile regression model in which the lifetime variable is a defective distribution.

Under the proposed approach, any defective distribution can be considered for the lifetime variable. Motivated by the work of[24] and by the real data features presented in the Subsection 1.2, we used the defective version of the generalized Gompertz distribution introduced by[40]. An important feature of the generalized Gompertz distribution is the different forms of the hazard function depending on the values of its parameters. This distribution is conveniently reparametrized in terms of the q-th quantile and then linked to covariates by means of a logarithm link function, which allow us to estimate the quantiles directly.

In this work, the estimation procedure is constructed under the Bayesian perspective of inference.[36] discuss several advantages of using the Bayesian approach, such as the possibility of making exact inferences for any sample size without the need for asymptotic calculations, the ease of estimating variances and any other measures of the posteriori distribution, and the use of a priori information. There are situations where information through expert opinion and/or past experience can be expressed in the a priori distributions. However, in a scenario where no prior information about the failure time is available, non-informative priors can be considered.

An important detail of our approach is that we use a parametric quantile regression (PQR) model, similarly to what[37] considered for their generalized gamma proposal, though in a classic framework. This concept was thoroughly discussed by[41], where the author argued that the idea to model the quantile function of a probability distribution, instead of another location parameter, for instance, adds much more flexibility to one's modelling strategy. Specifically in comparison to the usual quantile regression defined by[30], this technique presents a great advantage, as we are able to avoid a big concern about quantile regression models, which is the issue with crossing quantiles. Because we are considering the defective distribution generalized Gompertz to model these quantiles and correspondingly its cure fraction, we coin this model as a defective cure rate quantile regression model. We believe this proposal brings more adaptability to current modeling efforts of survival times.

1.2 | Motivating example: Male breast cancer diagnosed in the state of São Paulo

Breast cancer in women has been extensively studied, as well as its prognostic factors. The large number of studies is certainly comprehensible, as this type of cancer corresponds to almost 30% of new cancer cases diagnosed annually in Brazilian women, being a major cause of death in the female population (according to Brazilian National Institute of Cancer - www.inca.gov.br). In fact, this is the second most frequent cancer type that occurs in women, after non-melanoma skin neoplasms, but there is still some misunderstanding on its incidence among men as well. According to the Centers for Disease Control and Prevention (CDC - www.cdc.gov), about 1 out of every 100 breast cancers diagnosed is found in a man, however, the occurrence of this neoplasm tends to increase due to the poor quality of life and the difficulty in making an early diagnosis. In a survey of scientific articles on breast cancer in men,[1] pointed out that the incidence has increased significantly from 0.86 to 1.06 per 100,000 men over the past few decades; the highest rates occurring in North America and Europe and the lowest rates in Asia. Additionally, it was observed that men with breast cancer have the worst overall survival rates in relation to women, but this is probably due to their older age at the time of diagnosis, which corresponds to the most advanced stage of presentation of the disease, as well as higher death rates due to disease comorbidity.

In view of this problem, it is very desirable to establish which prognostic factors impact the survival of men with breast cancer, in order to understand how the disease works in this population. This information would allow physicians to prevent the overall progressive burden of this disease, with measures of control and preventive interventions in this context.

For that, we consider a male breast cancer dataset from a retrospective survey of 887 records of males diagnosed with breast in the state of São Paulo, Brazil, between 2000 and 2019, with follow-up conducted until February of 2020 and with at least two months of follow-up. Death due to breast cancer was defined as the event of interest. Those patients who did not die due to breast cancer during the follow-up period were characterized as right-censored observations. In this analysis, we want to study the impact of clinical stage, age group (<55; 55 - 65 and >65 years old) and different treatments (surgery, chemotherapy, radiotherapy and hormone therapy) in survival for men, in order to understand how the disease affects this population. The age groups were determined by mastologists from our work team.

This dataset is provided by the Fundação Oncocentro de São Paulo (FOSP), which is responsible for coordinating and monitoring the implementation of the Hospital Cancer Registry in the State of São Paulo (Brazil), in addition to systematizing and evaluating cancer care data available for the state. The FOSP is a public institution connected to the State Health Secretariat that monitors the evolution of the Oncology Care Network, assists the State Department of Health in the creation and application of prevention and health promotion programs, and monitors the evolution of cancer mortality in the state.

Of the 887 patients, 77% did not die during the follow-up period, that is, 685 patients have a right-censored event time. A total of 270 (30.4%) patients are under the age of 55, 280 (31.6%) are between 55 and 65 years old and 337 (38%) are older than 65 years old. Besides that, 21.1% of the patients are classified as clinical stage I, 37.2% as clinical stage II, 30.2% as clinical stage III and 11.5% as clinical stage IV. In relation to the cancer treatment, 69.1% underwent surgery, 41% had radiotherapy sessions, 62.9% did chemotherapy and 52.4% had in their treatment hormone therapy.

Figure 1 presents the Kaplan-Meier estimates grouped by each explanatory variable. Of note, there is a strong evidence of long-term survivors for all groups. Among all of the variables considered in our study, those with clinical stage I had a better prognosis, and stage IV had the worst one. Intermediate-age men (between 55 and 65 years old) have the highest survival rate when compared to younger and older patients, and these last two have very similar behavior. Patients who underwent surgery, did not need or did not undergo chemotherapy and underwent hormone therapy have the best prognosis. The behavior of those who did and did not undergo radiotherapy is similar.

Besides estimating the cure rate, physicians from our working group are interested in assessing how the regression coefficients of age, clinical stage and the different treatments behave at different quantiles of the survival time.

1.3 | Outline of the paper

We organize the rest of the paper as follows: the model formulation is described in Section 2. Parameter inference under a Bayesian perspective is discussed in Section 3. In Section 4, we carry out a Monte Carlo simulation study to demonstrate the performance of the proposed estimation method. An application to men with breast carcinoma is discussed in Section 5. Furthermore, some conclusions are mentioned in Section 6.

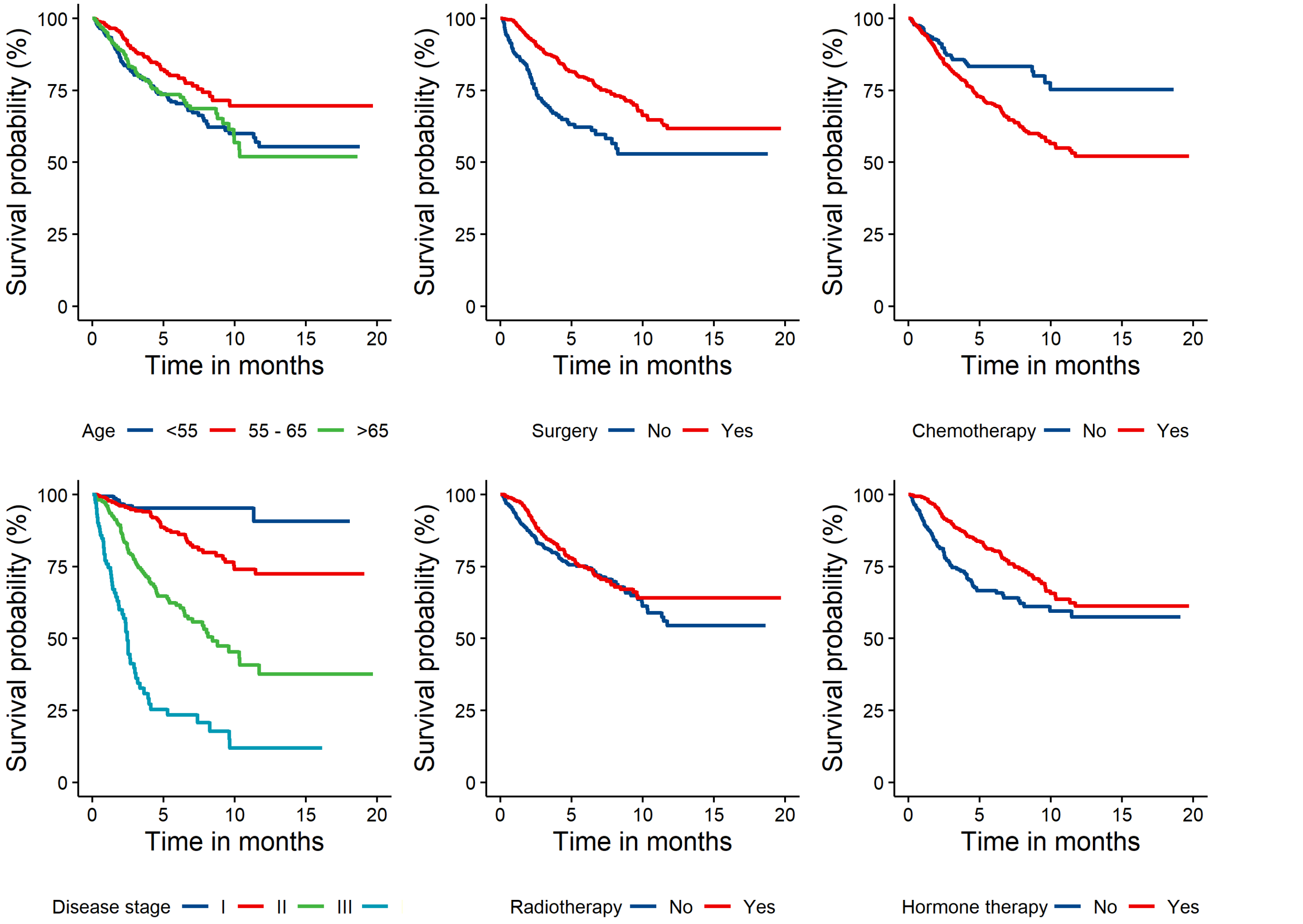

FIGURE 1 Kaplan-Meier estimates for the male breast cancer dataset grouped by age, clinical stage and different treatments.

2 | BACKGROUND

In this section we present the development of the proposed model. In Subsection 2.1 the generalized Gompertz distribution and its defective version are presented and the proposed reparametrization is presented in Subsection 2.2, resulting in the generalized Gompertz cure rate quantile regression model.

2.1 | Generalized Gompertz distribution and its defective version

The generalized Gompertz (GG) distribution introduced by[40] considers that lifetime T conditional to parameters λ, α and θ, has density and survival functions, respectively, given by

$$f(t \mid \lambda, \alpha, \theta) = \lambda\theta \exp\left\{\alpha t - \frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\left(1 - \exp\left\{-\frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\right)^{\theta-1}, \quad t > 0 \tag{1}$$

and

$$S(t \mid \lambda, \alpha, \theta) = 1 - \left(1 - \exp\left\{-\frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\right)^{\theta}, \tag{2}$$

where $\lambda > 0$ and $\alpha > 0$ are scale parameters and $\theta > 0$ is a shape parameter.

The GG distribution includes the following distributions as special cases: (i) generalized exponential distribution[42] when α tends to zero; (ii) Gompertz distribution when $\theta = 1$; and (iii) exponential distribution when α tends to zero and $\theta = 1$.

The hazard function of the GG distribution is given by

$$h(t \mid \lambda, \alpha, \theta) = \frac{\lambda\theta \exp\left\{\alpha t - \frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\left(1 - \exp\left\{-\frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\right)^{\theta-1}}{1 - \left(1 - \exp\left\{-\frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\right)^{\theta}}.$$

An important feature of the GG distribution is the different forms of the hazard function depending on the values of its parameters. The shape of hazard function of the GG distribution is[40]:

(i) increasing if $\alpha > 0$ and $\theta = 1$;

(ii) constant if $\alpha = 0$ and $\theta = 1$;

(iii) increasing when $\theta > 1$;

(iv) decreasing if $\alpha = 0$ and $\theta < 1$; and

(v) bathtub if $\alpha > 0$ and $\theta < 1$.

The GG distribution becomes a defective distribution, characterized by having a probability density function that integrates to values less than 1[11], if $\alpha < 0$; thus being suitable for modeling survival data with long-term survivors. The corresponding cure fraction is:

$$\lim_{t\to\infty} S(t \mid \lambda, \alpha, \theta) = 1 - \left[1 - \exp\left(\frac{\lambda}{\alpha}\right)\right]^{\theta} = p_0(\lambda, \alpha, \theta) \in [0, 1]. \tag{3}$$

Unlike standard mixture models, defective distributions have the advantage of allowing a cure rate without adding any extra parameters in the modeling Scudilio et al.[20].

As mentioned before, one of this work's goals is to estimate the quantile function. Nevertheless, the quantile function is not well defined, as the distribution is defective. To circumvent this restriction, we will rewrite the defective distribution as a mixture model, i.e., we will separate the susceptible and immune populations without including an additional parameter in the modeling; thus, in addition to ensuring the advantage of parsimony, we will also be able to describe the quantiles of the susceptible population.

2.2 | Generalized Gompertz Cure Rate Quantile Regression (GGCRQR)

To consider a quantile regression model, first we write the model in (2) as a standard mixture model[2,3]. In this case, we define an indicator variable Z, which takes the value 0 if the subject is immune and 1 if the subject is susceptible. Let $\mathbb{P}[Z = 0] = p_0(\lambda, \alpha, \theta)$ and $\mathbb{P}[Z = 1] = 1 - p_0(\lambda, \alpha, \theta)$, where $p_0(\lambda, \alpha, \theta)$ is given by (3).

Then, the survival function in (2) is given by

$$S(t \mid \lambda, \alpha, \theta) = p_0(\lambda, \alpha, \theta) + \left[1 - p_0(\lambda, \alpha, \theta)\right] S_1(t \mid \lambda, \alpha, \theta), \tag{4}$$

where $S_1(\cdot)$ denotes the survival function of the susceptibles. From (4), we can get an expression for S_1 as

$$S_1(t \mid \lambda, \alpha, \theta) = \frac{S(t \mid \lambda, \alpha, \theta) - p_0(\lambda, \alpha, \theta)}{1 - p_0(\lambda, \alpha, \theta)}. \tag{5}$$

By considering the expression in (5), we note that S_1 is a proper survival function with a quantile function given by

$$\mu_q = \mu(q \mid \lambda, \alpha, \theta) = \frac{\log\left(\lambda - \alpha \log\left\{1 - q^{\frac{1}{\theta}}\left[1 - \exp\left(\frac{\lambda}{\alpha}\right)\right]\right\}\right) - \log(\lambda)}{\alpha}, \quad 0 < q < 1. \tag{6}$$

Next, we will reparameterize the density function in (1) in terms of q-th quantile μ_q, such that θ can be written as

$$\theta = -\frac{\log(q)}{\log\left[1 - \exp\left(\frac{\lambda}{\alpha}\right)\right] - \log\left\{1 - \exp\left[-\frac{\lambda}{\alpha}\left(e^{\alpha\mu_q} - 1\right)\right]\right\}}, \tag{7}$$

where this result is obtained by isolating θ in Eq. (6).

It is important to highlight that as we write the survival function for susceptible individuals (Equation (5)) as a function of the survival function of the defective generalized Gompertz distribution and cure fraction $p_0(\lambda, \alpha, \theta)$, in which for them $\alpha < 0$, we have that the parametric space for α is $(-\infty, 0)$ in the reparametrized defective generalized Gompertz distribution.

If we replace (7) in (1), (2) and (3), we have the reparameterized density and survival functions and cure fraction, which are given by, respectively,

$$f(t \mid \lambda, \alpha, \mu_q) = \lambda \left(- \frac{\log(q)}{\log\left[1 - \exp\left(\frac{\lambda}{\alpha}\right)\right] - \log\left\{1 - \exp\left[-\frac{\lambda}{\alpha}\left(e^{\alpha\mu_q} - 1\right)\right]\right\}} \right)$$
$$\times \exp\left\{\alpha t - \frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}$$
$$\times \left(1 - \exp\left\{-\frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\right)^{-\frac{\log(q)}{\log\left[1-\exp\left(\frac{\lambda}{\alpha}\right)\right]-\log\left\{1-\exp\left[-\frac{\lambda}{\alpha}\left(e^{\alpha\mu_q}-1\right)\right]\right]\right\}}-1}, \tag{8}$$

$$S(t \mid \lambda, \alpha, \mu_q) = 1 - \left(1 - \exp\left\{-\frac{\lambda}{\alpha}\left[\exp(\alpha t) - 1\right]\right\}\right)^{-\frac{\log(q)}{\log\left[1-\exp\left(\frac{\lambda}{\alpha}\right)\right]-\log\left\{1-\exp\left[-\frac{\lambda}{\alpha}\left(e^{\alpha\mu_q}-1\right)\right]\right]\right\}}} \tag{9}$$

and

$$p_0(\lambda, \alpha, \mu_q) = 1 - \left[1 - \exp\left(\frac{\lambda}{\alpha}\right)\right]^{-\frac{\log(q)}{\log\left[1-\exp\left(\frac{\lambda}{\alpha}\right)\right]-\log\left\{1-\exp\left[-\frac{\lambda}{\alpha}\left(e^{\alpha\mu_q}-1\right)\right]\right]\right\}}}. \tag{10}$$

Hereafter, we shall use the notation $T \sim dGG(\alpha, \lambda, \mu_q, q)$ where $\mu_q > 0$ is the quantile parameter, $\lambda > 0$, $\alpha < 0$, and $q \in (0, 1)$ is known.

Now, we build the defective generalized Gompertz quantile regression model, imposing that the quantile μ_q of T satisfies the following functional relation:

$$\mu_q(\boldsymbol{\beta}_q, \mathbf{x}) = \mu_q = \exp(\mathbf{x}^\top \boldsymbol{\beta}_q), \tag{11}$$

in which $\mathbf{x}^\top = (1, x_1, \dots, x_p)$ is the vectors of covariates and $\boldsymbol{\beta}_q = (\beta_{q0}, \beta_{q1}, \dots, \beta_{qp})^\top$ is the unknown vector of regression parameters to be estimated.

The parameters α and λ can also be linked to covariates, which gives more flexibility to the model. In this way, for α, we have that

$$\alpha = -\exp(\mathbf{w}^\top \boldsymbol{\gamma}), \tag{12}$$

in which $\mathbf{w}^\top = (1, w_1, \dots, w_r)$ are the covariates vector and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_r)^\top$ is the unknown vector of parameters related to α.

Finally, for λ, we have

$$\lambda = \exp(\mathbf{z}^\top \boldsymbol{v}), \tag{13}$$

in which $\mathbf{z}^\top = (1, z_1, \dots, z_s)$ is the covariate vector and $\boldsymbol{v} = (v_0, v_1, \dots, v_s)^\top$ is the relative unknown vector of parameters related to λ.

After this formulation, we have that the parameter vector is $\boldsymbol{\vartheta}_q^1 = (\boldsymbol{\beta}_q^\top, \boldsymbol{\gamma}^\top, \boldsymbol{v}^\top)^\top$. In the next section, we discuss how to obtain parameter estimates for this vector. Thus, the proposed model under the generalized Gompertz distribution is referred to as Generalized Gompertz Cure Rate Quantile Regression (GGCRQR).

Our formulation is related to the approach also known as distributional regression[27,28], where all parameters of the conditional distribution are explained as function of covariates, through properly chosen link functions. Although we do something similar here, we are still interested in studying the effects of each covariate in the quantile function of the Generalized Gompertz distribution, while also analysing the conditional cure rate given by its defective version, which would put our model in a slightly different class.

3 | INFERENCE

Consider that the lifetime T is possibly not observed, that is, it is constrained by a right censored failure time and let C denote the censoring time. In a sample of size n, we then observe the i-th lifetime $t_i = \min\{T_i, C_i\}$, and i-th failure indicator $\delta_i = I(T_i \leq C_i)$, where $\delta_i = 1$ if T_i is observed and $\delta_i = 0$ otherwise, for $i = 1, \dots, n$.

We consider that T_i's are independent random variables with density and survival functions $S\left(t_i \mid \boldsymbol{\vartheta}_q\right)$ and $f\left(t_i \mid \boldsymbol{\vartheta}_q,\right)$, obtained by replacing $\mu_{qi} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_q)$, $\alpha_i = -\exp(\mathbf{w}_i^\top \boldsymbol{\gamma})$ and $\lambda_i = \exp(\mathbf{z}_i^\top \boldsymbol{v})$ in Equations (8) and (9), respectively, where $\mathbf{x}_i^\top = (1, x_{i1}, \dots, x_{ip})$, $\mathbf{w}_i^\top = (1, w_{i1}, \dots, w_{ir})$ and $\mathbf{z}_i^\top = (1, z_{i1}, \dots, z_{is})$, for $i = 1, \dots, n$. We assume that each censoring time C_i is independent of lifetime T_i, for all $i = 1, \dots, n$, and we consider a noninformative censoring assumption, i.e., the censoring distribution does not involve the parameters of the distribution of T. Therefore, the likelihood function of $\boldsymbol{\vartheta}_q$ can be written as

$$L(\boldsymbol{\vartheta}_q; \mathbf{D}) \propto \prod_{i=1}^{n} \left[f\left(t_i \mid \boldsymbol{\vartheta}_q\right)\right]^{\delta_i} \left[S\left(t_i \mid \boldsymbol{\vartheta}_q\right)\right]^{1-\delta_i},$$

where $\mathbf{D} = (n, \mathbf{t}, \boldsymbol{\delta}, X, W, Z)$, with $\mathbf{t} = \left(t_1, \ldots, t_n\right)^\top, \boldsymbol{\delta} = \left(\delta_1, \ldots, \delta_n\right)^\top$. Furthermore, $X = \left(\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top\right), W = \left(\mathbf{w}_1^\top, \ldots, \mathbf{w}_n^\top\right)$ and $Z = \left(\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top\right)$ are $n \times (p+1)$, $n \times (r+1)$ and $n \times (s+1)$ matrices containing the covariates information, respectively.

Under the Bayesian approach, we can specify the posterior distribution of $\boldsymbol{\vartheta}_q$ as

$$\pi(\boldsymbol{\vartheta}_q \mid \mathbf{D}) \quad \propto \quad \pi(\boldsymbol{\vartheta}_q) L(\boldsymbol{\vartheta}_q; \mathbf{D}), \tag{14}$$

where $\pi(\boldsymbol{\vartheta}_q)$ is the prior distribution of $\boldsymbol{\vartheta}_q$ for a fixed q.

The prior distributions for $\boldsymbol{\vartheta}_q$ can be defined in the following manner, assuming that they are prior independent. We assume a normal distribution for each β, γ and υ with mean 0 and variance 100. This setup will be considered in both the simulation and the application in the following sections.

The posterior distribution in (14) does not have a closed form and the parameters are estimated through simulated samples of the posterior distribution obtained by the Adaptive Metropolis algorithm with multivariate normal distribution as proposed by[43]. This approach is implemented in the statistical package *LaplacesDemon*[44], which provides a friendly environment for Bayesian inference within the R program[45].

As a result, a sample of size M from the joint posterior distribution of $\boldsymbol{\vartheta}_q$ is obtained (eliminating burn-in and jump samples). The sample from the posterior can be expressed as $(\boldsymbol{\vartheta}_q^{(1)}, \boldsymbol{\vartheta}_q^{(2)}, \ldots, \boldsymbol{\vartheta}_q^{(M)})$.

The posterior mean of $\boldsymbol{\vartheta}_q$, for instance, can be approximated by

$$\widehat{\boldsymbol{\vartheta}}_q = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\vartheta}_q^{(m)},$$

and the posterior mean of the cure rate is approximated by

$$\widehat{p}_0 = \frac{1}{M} \sum_{m=1}^{M} 1 - \left[1 - \exp\left(\frac{\lambda^{(m)}}{\alpha^{(m)}}\right)\right]^{-\frac{\log(q)}{\log\left[1-\exp\left(\frac{\lambda^{(m)}}{\alpha^{(m)}}\right)\right]-\log\left\{1-\exp\left[-\frac{\lambda^{(m)}}{\alpha^{(m)}}\left(e^{\alpha^{(m)}\mu_q^{(m)}}-1\right)\right]\right\}}}.$$

The Conditional Predictive Ordinate (CPO) for ith observation, CPO_i, is the probability to observe t_i when the model is fitted without this observation. Let $\mathbf{D}_{-i}$, the data without ith observation. The ith CPO is given by

$$CPO_i = \int\limits_{\Theta_q} f(t_i \mid \boldsymbol{\vartheta}_q) \pi(\boldsymbol{\vartheta}_q \mid \mathbf{D}_{-i}) d\boldsymbol{\vartheta}_q = \left\{ \int\limits_{\Theta_q} \frac{\pi(\boldsymbol{\vartheta}_q \mid \mathbf{D})}{f(t_i \mid \boldsymbol{\vartheta}_q)} d\boldsymbol{\vartheta}_q \right\}^{-1},$$

in which Θ_q is the parameter space of ϑ_q.

The CPO_i can be estimated by considering the M values of the posterior distribution of ϑ_q[46]:

$$\widehat{CPO}_i = \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{1}{g(t_i \mid \boldsymbol{\vartheta}_q^{(m)})} \right\}^{-1},$$

and $g(t_i \mid \boldsymbol{\vartheta}_q^{(m)})$ is $f(t_i \mid \boldsymbol{\vartheta}_q^{(m)})$ if $\delta_i = 1$ and $S(t_i \mid \boldsymbol{\vartheta}_q^{(m)})$ if $\delta_i = 0$, for $i = 1, \ldots, n$.

High values of CPO_i indicate that the model is able to describe the ith observation adequately and, for this reason, it is natural to think of a model selection measure that is a function of the CPO. The measure LPML (Log Pseudo Marginal Likelihood) is the sum of the logarithms of the CPO's of all observations, i.e., $LPML = \sum_{i=1}^{n} \log(\widehat{CPO}_i)$; the higher its value, the better will be the model fit.

4 | A SIMULATION STUDY

In this section, we evaluate the performance of the proposed model considering a simulation study. We only consider one predictor variable, as we believe adding more variables would not cause problems for the estimation algorithm, though this could be further investigated in the future.

For this simulation scenario, we consider the following values for the parameters: $\alpha = -0.25 = -\exp(\gamma_0) = -\exp(-1.386)$, $\lambda = 1 = \exp(\upsilon_0) = \exp(0)$, $\beta_{q0} = 1.3$, $\beta_{q1} = 0.7$ and $q \in \{0.2, 0.5, 0.8\}$. Moreover, six sample sizes are considered: $n =$ 50, 100, 300, 500, 1,000, 2,000. For each combination of parameter values and sample size, $B = 1{,}000$ datasets are generated by considering the algorithm 1.

Algorithm 1 Data generation algorithm.

```
1: Determine desired values for $\gamma_0$, $v_0$, $q$ and $\boldsymbol{\beta}_q = (\beta_{q0}, \beta_{q1})^\top$;
2: Define the proportion of censored data, given by $pc_0$ and $pc_1$, for $x = 0$ and $x = 1$, respectively;
3: For the $i$th subject, draw $x_i \sim$ Bernoulli(0.5), and calculate $p_{0x_i}$, in which $p_{00}$ and $p_{01}$ are given by (10) when $x_i = 0$ and $x_i = 1$, respectively;
4: Draw $u_i \sim$ Uniform(0, 1). If $u_i < p_{0x_i}$, set $w_i = \infty$; otherwise, generate $u_{1i} \sim U(0, 1 - p_{0x_i})$ and calculate
       $$w_i = \frac{1}{\alpha} \log \left[ 1 - \frac{\alpha}{\lambda} \log \left( 1 - u_{1i}^{(1/\theta_{x_i})} \right) \right], \text{ with } x_i = 1 \text{ or } x_i = 0; \qquad (15)$$
5: Draw $c_i \sim U(0, \tau_i)$, where $\tau_i$ is defined to have approximately $p_{cx_i}$ proportion of censoring data.
6: Determine $t_i = \min\{w_i, c_i\}$. If $t_i = w_i$, set $\delta_i = 1$, otherwise $\delta_i = 0$;
7: Repeat steps 3 to 6 for all $i = 1, \dots, n$. The data set for the $i$th subject is $\{t_i, x_i, \delta_i\}$, $i = 1, \dots, n$.
```

It is worth mentioning that, when $q = 0.2$ the cure fractions are $p_{00} = 0.32$ and $p_{01} = 0.83$; when $q = 0.5$ the cure fractions are $p_{00} = 0.15$ and $p_{01} = 0.54$ and, finally, when $q = 0.8$ the cure fractions are $p_{00} = 0.05$ and $p_{01} = 0.22$.

In order to obtain posterior quantities, the first 10,000 samples were discarded as burn-in samples. A jump of size 10 was chosen so that the correlation between the simulated values was close to zero. Thus, we get $M = 1,000$ simulated values for each parameter. We consider the posterior means for each parameter as our point estimates.

In Figure 2, we can check that the bias decreases as the sample sizes increases, since the point estimates do approximate their respective true values. In particular to parameters α and β_0, for smaller sample sizes this bias is considerably large, given their respective mean standard error, but these values are minimal for sample size equal to 300. Moreover, in general we have that the mean standard error is relatively large for small sample sizes, such as 50 and 100, but this number rapidly decreases as the sample sizes get larger as well. An important observation for this simulation study is that the cure fractions estimates present low bias values even for small sample sizes and for all q considered.

In Figure 3 we compare the coverage probability of 95% credible intervals for each parameter, considering HPD and equal tailed credible intervals. We can observe that the credible intervals for the cure fractions are mostly conservative, for $q = 0.2$ and $q = 0.5$. Additionally, for some parameters and for larger sample sizes, the coverage probability stays under the nominal value of 95%. This is the case for example for α and λ for $q = 0.2$. Though this is not ideal, these estimates still do not vary far from their nominal value.

5 | APPLICATION

About 1 out of every 100 breast cancers diagnosed is found in a man (www.cdc.gov). Given this unequal distribution, it is not surprising that there is still some misunderstanding on its incidence and prognostic factors among men. In this analysis, we want to study the impact of age, clinical stage and treatments (surgery, chemotherapy, radiotherapy and hormone therapy) in survival for men, in order to understand how the disease affects this population.

As already said in the Subsection 1.2, we consider a male breast cancer dataset from a retrospective survey of 887 records of males diagnosed with breast in the state of São Paulo, Brazil, between 2000 and 2019, with follow-up conducted until February of 2020 and with at least two months of follow-up. The dataset considered here and the routines used to estimate the models are available in the following link: https://github.com/brsantos/ggcrqr. The descriptive analyses can be seen at Subsection 1.2.

In this section, the GGCRQR model is fitted in two versions. In the first one (version 1), the covariates are included only in the quantile function μ_q by considering the relation in (11):

$$\mu_q = \exp(\beta_{q0} + \beta_{q1} \times \text{stage}_{II} + \beta_{q2} \times \text{stage}_{III} + \beta_{q3} \times \text{stage}_{IV} + \beta_{q4} \times \text{age}_{\geq 55\& \leq 65} + \beta_{q5} \times \text{age}_{>65} + \beta_{q6} \times \text{surgery} + \beta_{q7} \times \text{chemo} + \beta_{q8} \times \text{hormone} + \beta_{q9} \times \text{radio}), \qquad (16)$$

where we consider $q = \{0.10, \dots, 0.90\}$ and the coefficients β_{qi}, $i = 0, 1, \dots, 9$, also vary with q.

In the second version (version 2), in addition to considering covariates in the quantile function as in (16), we also consider covariates in α when considering the relation presented in (12). The idea to include covariates in α is to make the model more flexible, as discussed before.

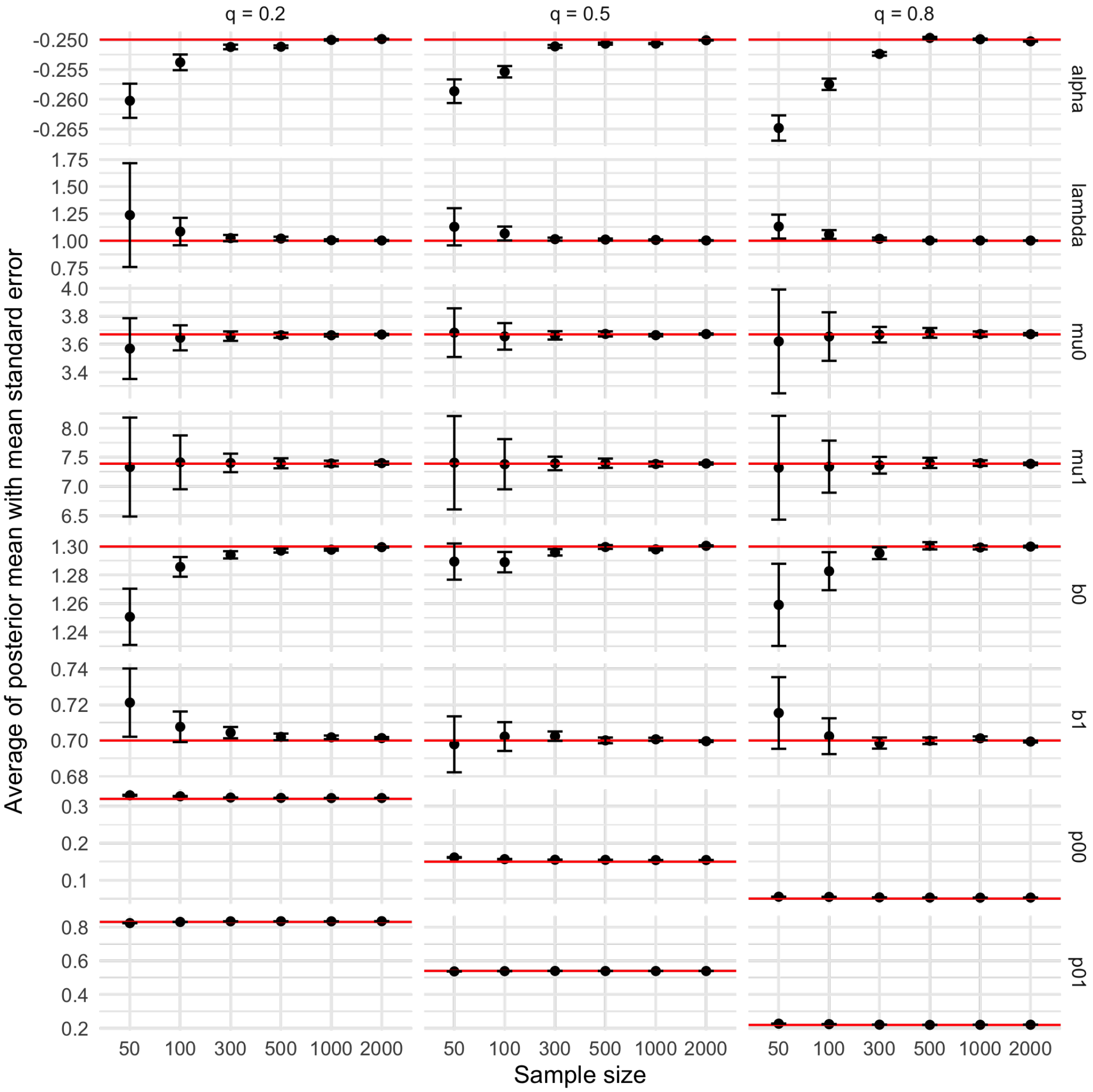

FIGURE 2 Posterior means for the different combinations of $q = \{0.2, 0.5, 0.8\}$, $n = 50,\ 100,\ 300,\ 500,\ 1{,}000,\ 2{,}000$ and the different parameters represented by the dots. The red line represents the true value for each parameter. The bar around the dot denotes the mean standard error.

For each version, the adaptive Metropolis-Hastings algorithm was run, discarding the first 50,000 iterations as burn-in samples and using a jump of size 100 to avoid correlation problems, with a sample size of $M = 2{,}000$. The convergence of the chain was evaluated by multiple runs of the algorithm from different starting values and was monitored through graphical analysis, where we were able to obtain good convergence results.

Figure 4 presents the coefficients values by considering q varying from 0.1 to 0.9 by 0.1 for the β_{q9} parameter in (16). The solid line represents the posterior mean value and dotted lines represent the 95% highest probability density (HPD) interval. There is no evidence for the quantiles of the survival times of susceptible individuals who underwent radiotherapy to be different from those who did not have radiotherapy, the reference category, for both versions of the GGCRQR model. Because of this, the two versions models are fitted again but without radiotherapy variable. Thus, the quantile function μ_q is given by:

$$\mu_q = \exp(\beta_{q0} + \beta_{q1} \times \text{stage}_{II} + \beta_{q2} \times \text{stage}_{III} + \beta_{q3} \times \text{stage}_{IV} + \beta_{q4} \times \text{age}_{\geq 55 \& \leq 65} + \beta_{q5} \times \text{age}_{>65} + \beta_{q6} \times \text{surgery} + \beta_{q7} \times \text{chemo} + \beta_{q8} \times \text{hormone}). \tag{17}$$

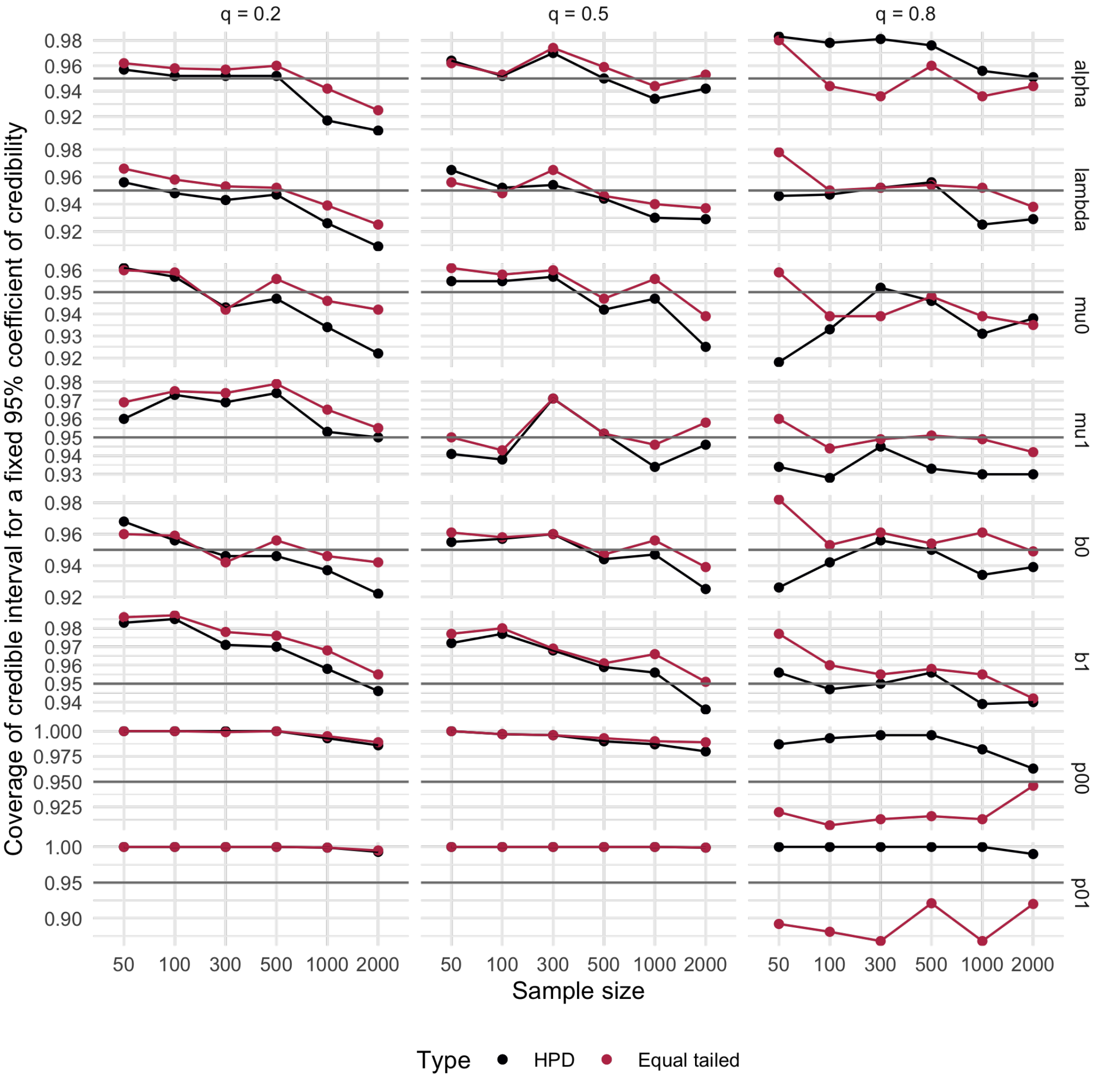

FIGURE 3 Coverage probability for the different combinations for $q = \{0.2, 0.5, 0.8\}$, $n = 50,\ 100,\ 300,\ 500,\ 1000,\ 2000$ and the different parameters. The gray line represents the nominal value of 0.95.

For the version with covariates in α, the selected model that presents the highest LPML is that with only surgery and hormone therapy, this is

$$\alpha = -\exp(\gamma_0 + \gamma_1 \times \text{surgery} + \gamma_2 \times \text{hormone}). \tag{18}$$

Figures 5 and 6 present the coefficients values by considering q varying from 0.1 to 0.9 by 0.1 for the parameters in Equation (17) for the model with covariates only in the quantile function and for the model with covariates in the quantile function and in the α parameter, as in Equation (18), respectively. The solid line represents the posterior mean value and dotted lines represent the 95% HPD interval. It is important to note how some coefficients vary along the different quantiles. For both GGCRQR model versions, all the coefficients HPD intervals do not include the zero value for all values of q considered.

Table 1 presents posterior mean and HPD 95% credible intervals for parameters related to α and λ, by considering the relations in Equations (12) and (13), respectively, for models versions 1 and 2. Since we have the posterior draws for each parameter we can apply the function in (10) for each draw so we have a posterior sample from the cure fraction. This allows us to check the posterior density for the combination of clinical stage and other covariates regarding its cure fraction as well. These densities are shown in Figures 7 and 8 for the model version 1 and for the model version 2, respectively.

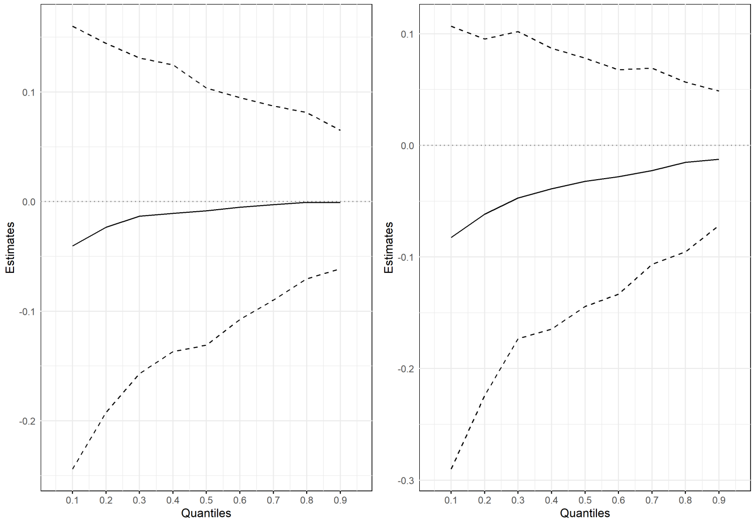

FIGURE 4 Posterior estimates for the regression parameter β_{q9} given the quantiles, q, from Equation (16) for version 1 (in the left) and for version 2 (in the right). The solid lines denote the posterior mean, while the dashed are the respective HPD credible interval.

The LPML values for each model version by considering different q values are presented in Table 2. As one can see, the version 2 (model with covariates in μ_q and in α) presents the highest LPML value for all q values. Because of this, this is the chosen GGCRQR model and its results are interpreted in the following.

TABLE 1 Posterior estimates for the parameters involved in α and λ parameters (they are the same for all q values).

	Version 1 - with covariates only in μ_q		Version 2 - with covariates in μ_q and α	
Parameter	Posterior mean	95% HPD credible interval	Posterior mean	95% HPD credible interval
γ_0	-1.604	-1.923 -1.322	-1.117	-1.487 -0.751
γ_1	-	-	-0.194	-0.462 -0.079
γ_2	-	-	-0.435	-0.701 -0.165
υ_0	-1.147	-1.516 -0.812	-1.017	-1.378 -0.614

TABLE 2 LPML for model versions 1 and 2 for all q values considered.

q	Version 1 - with covariates only in μ_q	Version 2 - with covariates in μ_q and α
0.1	-708.82	-705.44
0.2	-708.72	-705.21
0.3	-708.88	-704.45
0.4	-708.53	-704.22
0.5	-709.18	-704.07
0.6	-708.38	-703.88
0.7	-709.06	-703.73
0.8	-709.08	-703.98
0.9	-708.72	-704.10

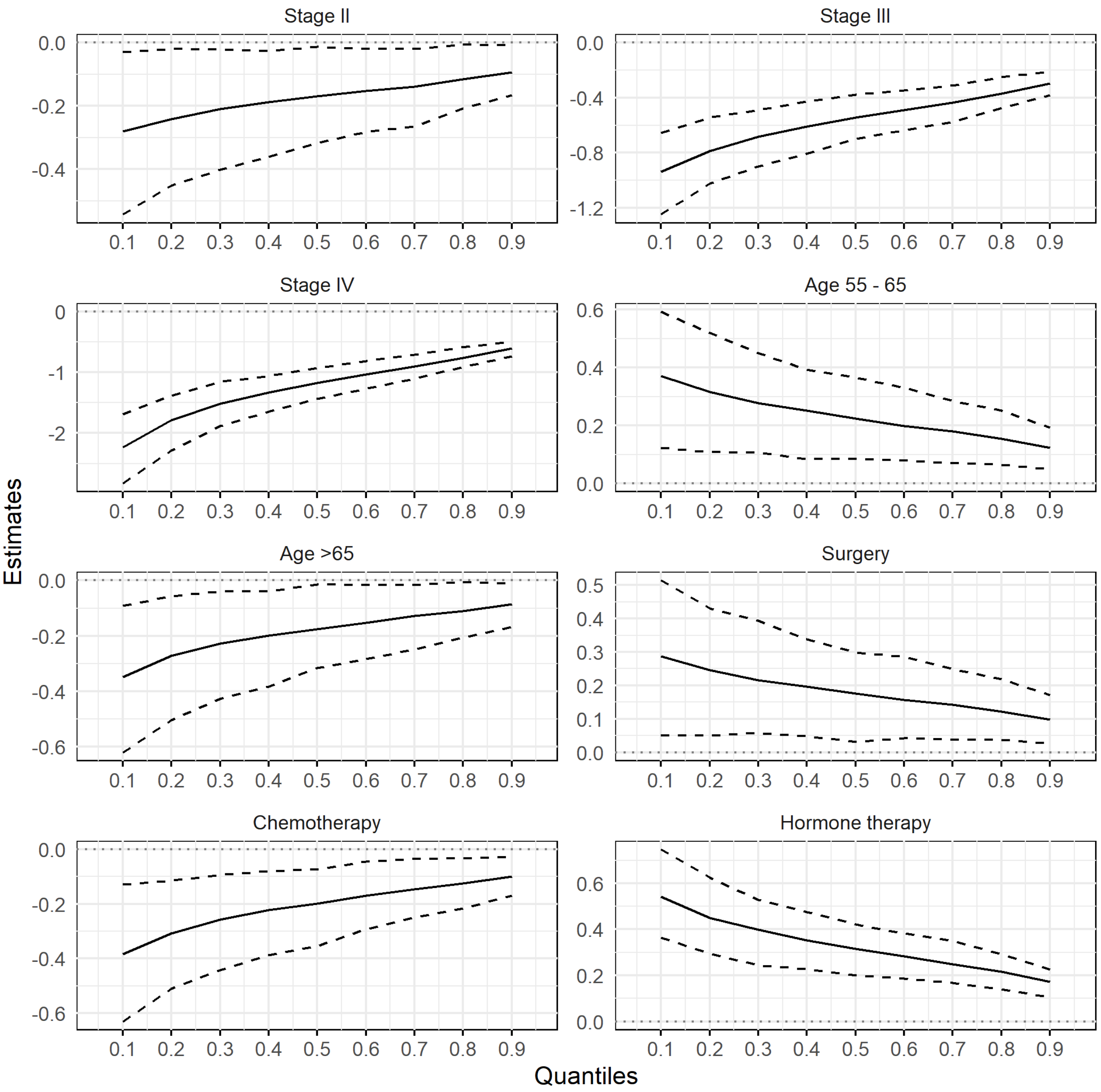

FIGURE 5 Posterior estimates for the regression parameter $\boldsymbol{\beta}$ given the quantiles, q, presented in Equation (17), with covariates only in the quantile function - version 1. The solid lines denote the posterior mean, while the dashed are the respective HPD credible interval.

By analysing the Figure 6, it is possible to note that the quantile-based model defined in (17) is able to estimate the effects for different quantiles and display contrasting differences along these quantiles. For clinical stage variable, the effect is increasingly negative, that is, the life time is shorter as the stage of the disease increases, with stage I as the reference category. Besides, men aged between 55 and 65 years old have the highest survival time and men aged over 65 years old have the lowest survival time. In relation to cancer treatments, those men who did surgery present the highest lifetime as well as those who underwent hormone therapy. On the other hand, those men undergoing chemotherapy have a shorter survival time.

In relation to the cure fraction (Figure 8, one can observe the ordering of the probability of cure for the different clinical stages is consistent with the common prognosis of this carcinoma, as people in Stage I have a higher probability of being cured, and men at Stage IV of the cancer diagnosis have a lower probability of being cured in comparison to other stages. Fixed the clinical stage, men older than 55 and younger than 65 years old have a higher probability of cure, in comparison with the other groups,

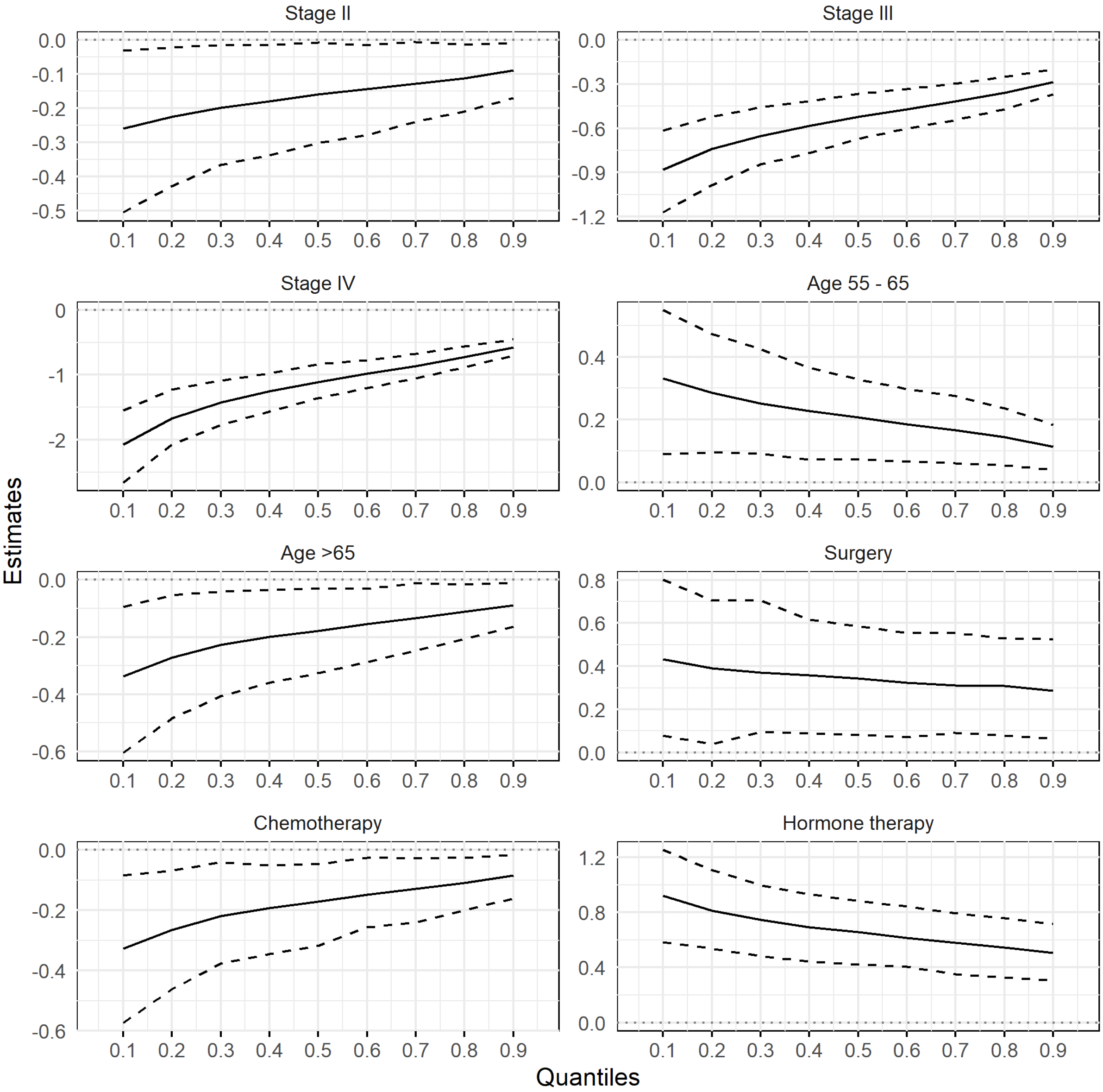

FIGURE 6 Posterior estimates for the regression parameter $\boldsymbol{\beta}$ given the quantiles, q, presented in Equation (17), with covariates in the quantile function and in α as in Equation (18) - version 2. The solid lines denote the posterior mean, while the dashed are the respective HPD credible interval.

and men aged over 65 years old have the lowest cure fraction, as well as those undergoing chemotherapy. It seems that whether or not to have surgery and whether or not to perform hormone therapy does not make much difference in the cure fraction, which confirms what can be observed in Kaplan Meier's estimates in Figure 1.

We observed that those men who underwent chemotherapy have the lowest survival. However, this result needs to be analyzed with caution. Individuals who undergo chemotherapy generally have the disease more severely.

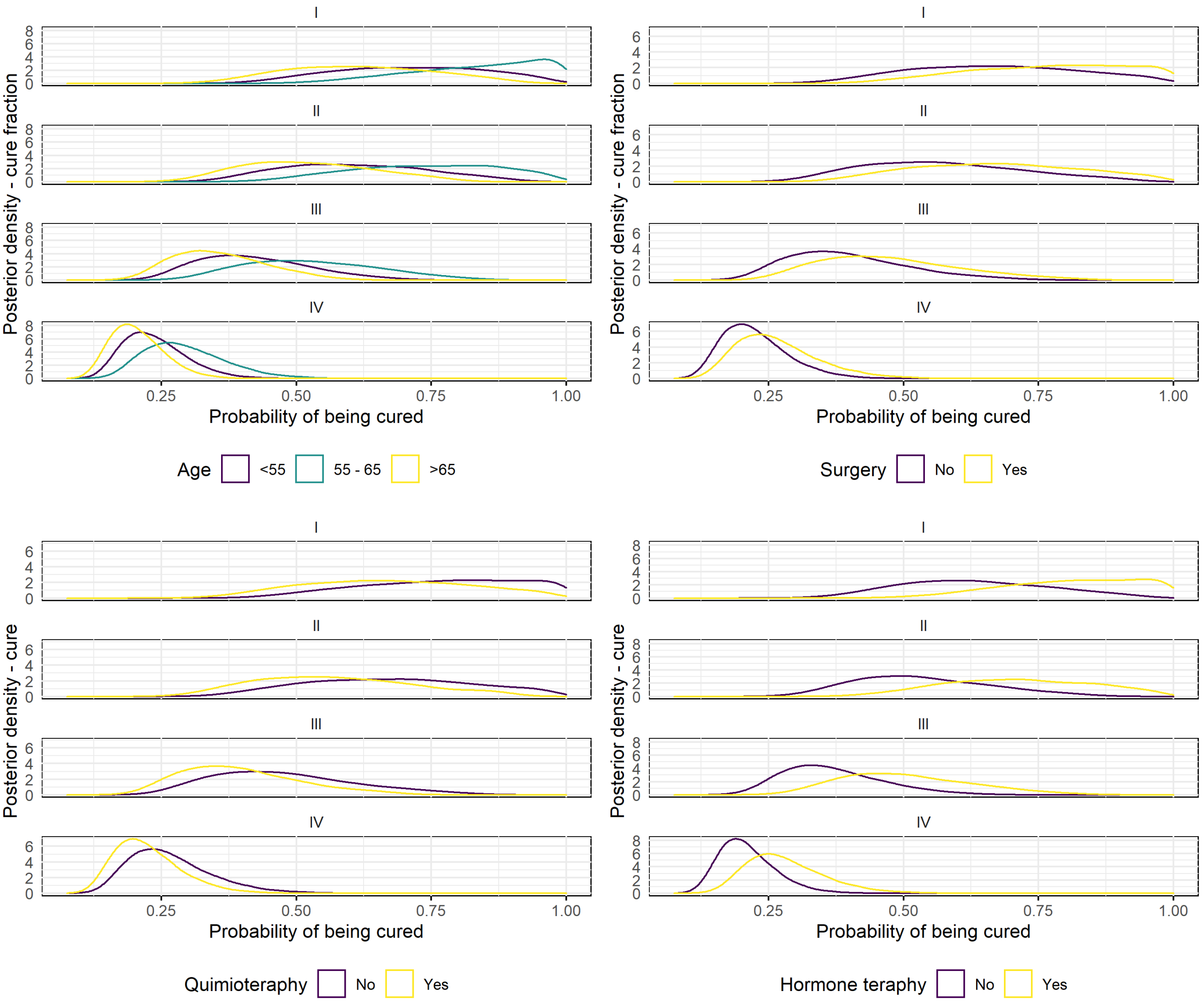

FIGURE 7 Posterior densities for the cure fraction parameter for each combination of clinical stage and other covariates by considering model version 1.

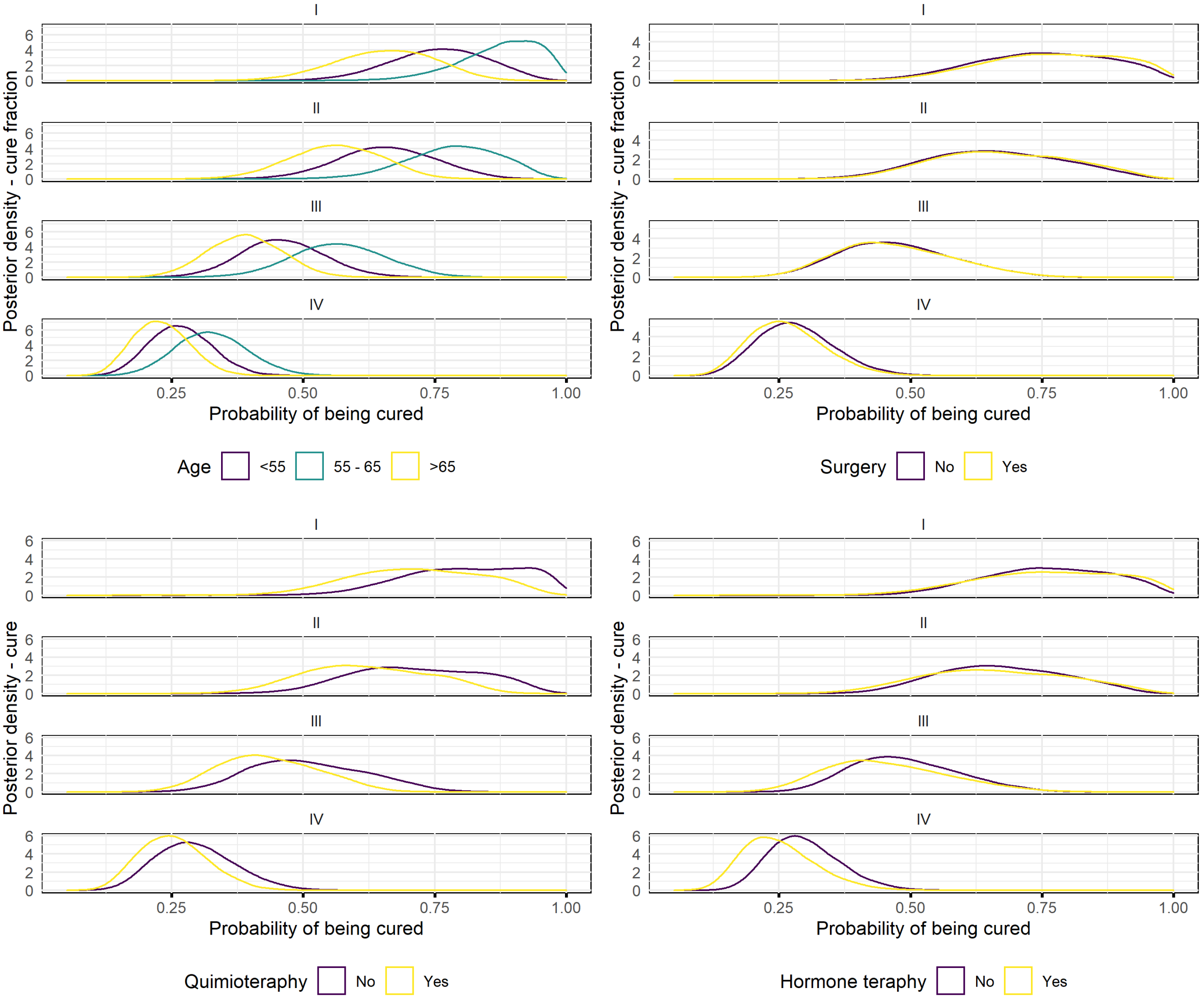

FIGURE 8 Posterior densities for the cure fraction parameter for each combination of clinical stage and other covariates by considering model version 2.

6 | CONCLUDING REMARKS

While the number of studies for female breast cancer is high, data about this carcinoma on the male population is still very limited[47]. Here we provide more information on this disease in the Brazilian population of men, where we discuss the association of clinical stage, age and treatments on the survival times of men enduring this type of cancer. Motivated by a demand from oncologists on the project team, the interest was to assess the impact of prognostic factors on different quantiles of survival time when cure rate is present. Our approach considered the use of a parametric quantile regression model, where we can observe how these variables affect the different quantiles of the survival times. By selecting a defective distribution in the form of the Generalized Gompertz distribution we are able to explain simultaneously both the quantiles of the survival times and the cure fraction, given the covariates, without adding new parameters to the probability distribution.

We have shown that our estimation approach based on MCMC draws effectively reach the correct values in a simulation study, while also presenting its main advantages in the application. Here we discussed how the quantile parameterization specifies a more complete picture of the conditional distribution of survival times, as we can study how each variable influences the different quantiles of this distribution. The parametric quantile regression approach also avoids the issue of quantile crossing, which is a common problem discussed in the literature for quantile regression models.

ACKNOWLEDGEMENTS

The authors thank the Fundação Oncocentro de São Paulo for providing the male breast cancer dataset.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

References

1. Haas P, Costa AB, Souza AP. Epidemiology of breast cancer in men. *Adolfo Lutz Institute Journal*. 2009;68:476–481.

2. Boag J. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society , Series B*. 1949;11(1):15-53.

3. Berkson J, Gage R. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*. 1952;42:501-515.

4. Farewell V. A model for binary variable with time-censored observations. *Biometrika*. 1977;64:43-46.

5. Farewell V. The use mixture models for the analysis of survival data with long term survivors. *Biometrics*. 1982;38:1041-1046.

6. Maller R, Zhou S. *Survival Analysis with Long Term Survivors*, Wiley, New York; 1996.

7. Banerjee S, Carlin BP. Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*. 2004;60:268-275.

8. Schnell P, Bandyopadhyay D, Reich BJ, Nunn M. A marginal cure rate proportional hazards model for spatial survival data. *Journal of the Royal Statistical Society, Series C: Applied Statistics*. 2015;64:673-691.

9. Yu B, Tiwari RC. A bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data. *The Canadian Journal of Statistics*. 2012;40:40-54.

10. Cox D, Oakes D. *Analysis of Survival Data*. London, Chapman & Hall; 1984.

11. Balka J, Desmond AF, McNicholas PD. Review and implementation of cure models based on first hitting times for wiener processes, *Lifetime Data Analysis*. 2009; 15:147-176.

12. Yakovlev, AY, Tsodikov, AD. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore; 1996.

13. Chen MH, Ibrahim JG, Sinha D. A new bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association*. 1999;94:909-914.

14. Yin G, Ibrahim JG. A general class of bayesian survival models with zero and nonzero cure fractions. *Biometrics*. 2005;61:403-412.

15. Rodrigues J, de Castro M, Balakrishnan N, Cancho VG. Destructive weighted poisson cure rate models. *Lifetime Data Analysis*. 2011;17: 333-346.

16. Rodrigues J, Cancho VG, de Castro M, Balakrishnan N. A bayesian destructive weighted poisson cure rate model and an application to a cutaneous melanoma data. *Statistical Methods in Medical Research*. 2012;21:585-597.

17. Klebanov LB, Rachev ST, Yakovlev A, Balakrishnan N. stochastic model of radiation carcinogenesis: Latent time distributions and their properties. *Mathematical Biosciences*. 1993;113:51-75.

18. Borges P, Rodrigues J, Balakrishnan N. Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data. *Statistics and Data Analysis*. 2012;56:1703-1713.

19. Borges P, Rodrigues J, Louzada F, Balakrishnan N. A cure rate survival model under a hybrid latent activation scheme. *Statistical Methods in Medical Research*. 2016;25:838-856.

20. Scudilio J, Calsavara VF, Rocha R, Louzada-Neto F, Tomazella V, Rodrigues AS. Defective models induced by gamma frailty term for survival data with cured fraction. *Journal of Applied Statistics*. 2019;46:484-507.

21. Cantor, AB, Shuster JJ. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*. 1992;11:931-937.

22. Gieser PW, Chang MN, Rao PV, Shuster JJ, Pullen J. Modelling cure rates using the Gompertz model with covariate information. *Statistics in Medicine*. 1998;17:831-839.

23. Rocha R, Nadarajah S, Tomazella V, Louzada F. Two new defective distributions based on the marshall aolkin extension. *Lifetime Data Analysis*. 2016;22:216-240.

24. Borges P. Em algorithm-based likelihood estimation for a generalized gompertz regression model in presence of survival data with long-term survivors: an application to uterine cervical cancer data, *Journal of Statistical Computation and Simulation*; 2017;87:1-11.

25. Martinez EZ, Achcar JA. A new straightforward defective distribution for survival analysis in the presence of a cure fraction. *Journal of Statistical Theory and Practice*. 2018;12:688-703.

26. Calsavara V, Rodrigues A, Tomazella V, Louzada F. Defective regression models for cure rate modeling with interval-censored data. *Biometrical Journal*. 2019;61:841-859.

27. Umlauf N, Klein N, Zeileis, A. BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). *Journal of Computational and Graphical Statistics*. 2018; 27:612-627.

28. Kneib T, Silbersdorff A, Säfken, B. Rage Against the Mean – A Review of Distributional Regression Approaches. *Econometrics and Statistics*. 2021.

29. Gupta C, Cobre J, Polpo A, Sinha D. Semiparametric Bayesian estimation of quantile function for breast cancer survival data with cured fraction. *Biometrical Journal*. 2016;58:1164-1177.

30. Koenker R, Bassett G. Regression quantiles. *Econometrica*. 1998;46:33-50.

31. Ying Z, Jung SH, Wei LJ. Survival Analysis with Median Regression Models. *Journal of the American Statistical Association*. 1995;90:178-184.

32. Honoré B, Khan S, Powell JL. Quantile Regression Under Random Censoring. *Journal of Econometrics*. 2002;109:67-105.

33. Portnoy S. Censored Regression Quantiles. *Journal of the American Statistical Association*. 2003;98:1001-1012.

34. Yin G, Zeng D, Li H. Power-Transformed Linear Quantile Regression With Censored Data. *Journal of the American Statistical Association*. 2008;103:1214-1224.

35. Peng L, Huang Y. Survival Analysis With Quantile Regression Models. *Journal of the American Statistical Association*. 2008;103:637-649.

36. Ibrahim JG, Chen MH, Sinha D. *Bayesian survival analysis* New York: Springer, 2001.

37. Noufaily A, Jones MC. Parametric quantile regression based on the generalized gamma distribution. *Journal of Royal Statistical Society, Series C: Applied Statistics*. 2013;62:723-740.

38. Xue X, Xie X, Strickler HD. A censored quantile regression approach for the analysis of time to event data. *Statistical Methods in Medical Research*. 2018;27:955-965.

39. Wu Y, Yin G. Cure rate quantile regression for censored data with a survival fraction. *Journal of the American Statistical Association*. 2013;108:1517-1531.

40. El-Gohary A, Alshamrani A, Al-Otaibi AN. The generalized Gompertz distribution. *Applied Mathematical Modelling*. 2012;37:13-24.

41. Gilchrist W. Regression Revisited. *International Statistical Review*. 2008;76:401-418.

42. Gupta RD, Kundu D. Generalized exponential distribution. *Australian and New Zealand Journal of Statistics*. 1999;41:173-188.

43. Haario, H. and Saksman, E. and Tamminen, J., Componentwise adaptation for high dimensional MCMC, *Computational Statistics* 20 (2005) 265–274.

44. Hall B. LaplacesDemon: An R Package for Bayesian Inference. *LaplacesDemon package manual*. 2012.

45. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. URL https://www.R-project.org/. 2020.

46. Gilks, WR and Richardson, S and Spiegelhalter, D. *Markov chain Monte Carlo in practice*, CRC press; 1995.

47. Yalaza M, Inan A, Bozer M. Male Breast Cancer. *Journal of Breast Health*. 2016;12:1-8.

□