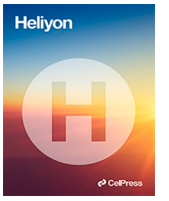
Heliyon 10 (2024) e32038

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A generalized Gompertz promotion time cure model and its fitness to cancer data

Ayesha Tahira, Muhammad Yameen Danish *

Department of Statistics, AIOU, Islamabad, Pin 44000, Pakistan

ARTICLE INFO

MSC:
62N01
62N05
62F10
62F15
62J05

Keywords:
Cure fraction
Promotion time cure model
Semiparametric maximum likelihood method
Generalized Gompertz distribution

ABSTRACT

The cure models based on standard distributions like exponential, Weibull, lognormal, Gompertz, gamma, are often used to analyze survival data from cancer clinical trials with long-term survivors. Sometimes, the data is simple, and the standard cure models fit them very well, however, most often the data are complex and the standard cure models don't fit them reasonably well. In this article, we offer a novel generalized Gompertz promotion time cure model and illustrate its fitness to gastric cancer data by three different methods. The generalized Gompertz distribution is as simple as the generalized Weibull distribution and is not computationally as intensive as the generalized F distribution. One detailed real data application is provided for illustration and comparison purposes.

1. Introduction

There are two main classes of cure rate models proposed in statistical literature, namely, mixture cure models [1–3] and promotion time cure models [4,5]. A broad class of flexible cure models including the promotion time cure model (PTCM) as special case is proposed in Ref. [6]. An extended parametric PTCM to incorporate the background mortality is developed in Ref. [7]. The two groups of mixture cure models and PTCMs are unified into one by introducing an extra transformation parameter to survival function using Box-Cox transformation [8–10]. The readers are referred to Refs. [11–17] for other extensions of the cure models, the corresponding baseline distributions, and the estimation methods. Recent advances in the developments and applications of the cure rate models can be seen in Refs. [18–24]. In the analysis of failure time data with cure fraction, several distributions including exponential, gamma, Weibull, lognormal, log-logistic, Burr type XII, Gompertz are commonly used as the baseline distributions. In this article, we study the generalized Gompertz distribution in promotion time cure model and compare its suitability as a baseline distribution with other well-known generalized distributions. The generalized Gompertz distribution was introduced by El-Gohary et al. [25] and has been used in different applications [26–29]. It can assume increasing, constant, decreasing or bathtub curve shapes for different combination of its parameter values unlike the Gompertz distribution which can have only monotonic increasing hazard. This is the main advantage of the generalized Gompertz distribution over the Gompertz distribution. Our main objectives in this paper are to study the PTCM with Gompertz and generalized Gompertz as the baseline distributions, and to compare these models with the PTCM based on Weibull, gamma, generalized Weibull and generalized gamma as the baseline distributions.

Presentation of the remaining paper is structured as follows. We discuss the PTCM and the distributions to be used as baseline in

* Corresponding author.
E-mail address: yameen.danish@aiou.edu.pk (M.Y. Danish).

https://doi.org/10.1016/j.heliyon.2024.e32038
Received 3 August 2023; Received in revised form 27 May 2024; Accepted 27 May 2024
Available online 1 June 2024

Section 2. Section 3 provides the nonparametric estimation of the survival function and then in Section 4, we consider maximum likelihood (ML) method of estimation and the associated inference. The procedure of semiparametric maximum likelihood estimation is provided in Section 5. Our main part of the paper belongs to results and discussions regarding the real data analysis, and it is presented in Section 6. The final Section 7 pertains to conclusion of the paper.

2. The model and its assumptions

Suppose that for a subject in a group of patients, the count of carcinogenic cells left active after some medication is a random variable, say N, distributed as Poisson with parameter θ. Further assume that for the ith cell, Z_i represent the random time taken to create a cancer disease for $i = 1, 2, 3, \ldots$, according to some probability distribution $F_Z(.)$ independently of the variable N. Then $T = \min\{Z_i, 0 \le i \le N\}$ with $P(Z_0 = \infty) = 1$ can be considered as the time to recurrence of cancer for the subject in question, where Z_0 denotes the time when there is no cancer. The PTCM is derived in Ref. [5] as

$$S_p(t) = \exp(-\theta + \theta S_Z(t)) = \exp(-\theta F_Z(t)) \tag{1}$$

with density and hazard functions as $f_p(t) = \theta f_Z(t)\exp(-\theta F_Z(t))$ and $h_p(t) = \theta f_Z(t)$, where $S_Z(.)$ denotes the survival function for the non-censored patients and $S_p(.)$ the survival function for the whole population. One can note that $S_p(\infty) = \exp(-\theta)$ represents the proportion of patients that are cured in the study.

We consider the generalized Gompertz as the baseline distribution for latency part of the PTCM to analyze the lifetimes of patients under risk of cancer disease and a log link function for incidence part of the PTCM to analyze the covariates affecting the cure rate of insusceptible patients. The generalized Gompertz distribution with parameters α, λ and δ is defined as

$$f_Z(z;\ \gamma) = \delta\alpha e^{\lambda z}\ e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\ \left(1 - e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\right)^{\delta-1};\ z \ge 0, \lambda \ge 0, \delta, \alpha > 0, \tag{2}$$

with distribution function as

$$F_Z(z;\ \gamma) = \left(1 - e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\right)^{\delta} \tag{3}$$

and hazard function as

$$h(z;\ \gamma) = \frac{\delta\alpha e^{\lambda z}\ e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\ \left(1 - e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\right)^{\delta-1}}{1 - \left(1 - e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\right)^{\delta}},$$

where γ denote the distribution parameters. We link the cure rate parameter θ to the covariate vector X as $\ln(\theta(X)) = X'\beta$, where β denote the vector of regression parameters. The cure fraction can be obtained from $p = \exp(-\exp(X'\beta))$. When we have only one covariate with values zero and one representing the group identification, then $p_0 = \exp(-\exp(B_0))$ and $p_1 = \exp(-\exp(B_0 + B_1))$ provide the cure fractions in the respective groups. Substituting the distribution function from (3) and $\ln(\theta(X)) = X'\beta$ in (1), we have the population survival function given by

$$S_p(t;\gamma,\beta|X) = \exp\left(-e^{X'\beta}\left(1 - e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\right)^{\delta}\right). \tag{4}$$

The associated density function from (2) and (3) is

$$f_p(t;\gamma,\beta|X) = \alpha\delta e^{X'\beta} e^{\lambda t} e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}$$

$$\times\left(1 - e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\right)^{\delta-1} \exp\left(-e^{X'\beta}\left(1 - e^{-\frac{\alpha}{\lambda}\left(e^{\lambda z}-1\right)}\right)^{\delta}\right). \tag{5}$$

Secondly, we consider the generalized Weibull distribution with density function

$$f_Z(z;\ \alpha,\lambda,\delta) = \alpha\lambda\delta z^{\alpha-1} e^{-\lambda z^{\alpha}} (1 - \exp(-\lambda z^{\alpha}))^{\delta-1};\ z \ge 0, \alpha > 0, \lambda > 0, \delta > 0, \tag{6}$$

and associated distribution function

$$F_Z(z;\ \alpha,\lambda,\delta) = (1 - \exp(-\lambda z^{\alpha}))^{\delta}. \tag{7}$$

The hazard function of the generalized Weibull distribution is

$$h(z;\ \alpha,\lambda,\delta) = \frac{\alpha\lambda\delta z^{\alpha-1}\ \exp(-\lambda z^{\alpha})\ (1 - \exp(-\lambda z^{\alpha}))^{\delta-1}}{1 - (1 - \exp(-\lambda z^{\alpha}))^{\delta}}.$$

We note that the generalized Weibull distribution with density in (6) reduces to standard Weibull distribution for $\delta = 1$ and can have different hazard shapes for different combinations of its parameter values. Further properties of the generalized Weibull distribution can be seen in Ref. [30]. For the generalized Weibull promotion time cure model, the survival function can be obtained from (1), and (7) with $\theta(X) = \exp(X'\beta)$ as

$$S_p(t;\alpha,\lambda,\delta,\beta|X)=\exp(-\exp(X'\beta)\,(1-\exp(-\lambda z^{\alpha}))^{\delta}), \tag{8}$$

and the corresponding population density function as

$$f_p(t;\alpha,\lambda,\delta,\beta|X)=\alpha\lambda\delta z^{\alpha-1}\exp(-\lambda z^{\alpha})\exp(X'\beta)$$

$$\times(1-\exp(-\lambda z^{\alpha}))^{\delta-1}\exp(-\exp(X'\beta)\,(1-\exp(-\lambda z^{\alpha}))^{\delta}). \tag{9}$$

Thirdly, we consider the generalized gamma distribution with density function

$$f_Z(z;\,\alpha,\lambda,\delta)=\frac{\alpha}{\Gamma(\delta)}\frac{z^{\alpha\delta-1}}{\lambda^{\alpha\delta}}\exp\left(-\left(\frac{z}{\lambda}\right)^{\alpha}\right);\quad z>0,\alpha>0,\lambda>0,\delta>0.$$

We can see that the generalized gamma distribution in (10) reduces to standard gamma distribution for $\alpha=1$ and standard Weibull distribution for $\delta=1$. Furthermore, as δ tends to infinity, the generalized gamma distribution tends to a lognormal distribution [31, 32]. One negative point for the generalized gamma distribution is that its distribution function does not exist in close form, one needs to perform numerical integration for

$$F_Z(z;\,\alpha,\lambda,\delta)=\int_0^z\frac{\alpha}{\Gamma(\delta)}\frac{u^{\alpha\delta-1}}{\lambda^{\alpha\delta}}\exp\left(-\left(\frac{u}{\lambda}\right)^{\alpha}\right).$$

The hazard function of the generalized gamma distribution is

$$h(z;\,\alpha,\lambda,\delta)=\frac{1}{1-F_Z(z;\,\alpha,\lambda,\delta)}\frac{\alpha}{\Gamma(\delta)}\frac{z^{\alpha\delta-1}}{\lambda^{\alpha\delta}}\exp\left(-\left(\frac{z}{\lambda}\right)^{\alpha}\right).$$

The population survival function is

$$S_p(t;\alpha,\lambda,\delta,\beta|X)=\exp(-\exp(X'\beta)\,F_Z(z;\,\alpha,\lambda,\delta)).$$

The corresponding population density function can be obtained from (12) as

$$f_p(t;\alpha,\lambda,\delta,\beta|X)=e^{X'\beta}\frac{\alpha}{\Gamma(\delta)}\frac{z^{\alpha\delta-1}}{\lambda^{\alpha\delta}}\exp\left(-\left(\frac{z}{\lambda}\right)^{\alpha}\right)e^{-\exp(X'\beta)\,F_Z(z;\,\alpha,\lambda,\delta)}.$$

3. Nonparametric estimation

Survival analysis can be carried out via three different methods, namely, parametric, nonparametric, and semiparametric. Each method has its own intuitive basis. While parametric methods are based on distribution assumptions relating to data generating mechanisms, nonparametric methods do not rely much on distribution assumptions. Semiparametric methods are considered better in the sense that they combine the efficiency of parametric methods and robustness of nonparametric methods. Nevertheless, we consider all three methods in this paper for comparison purposes.

Let the random variables $T_1,\ldots,T_n$ with the distribution function $F_T(t)$ and the density function $f_T(t)$ denote the failure times and the random variables $C_1,\ldots,C_n$ with the distribution function $F_C(c)$ and the density function $f_C(c)$ denote the censoring times for a random sample of n subjects entered into a life testing. Each patient in the sample will have either a failure time or censoring time. Let us denote the observed survival times as $Y_i=\min(T_i,C_i)$ and a censoring indicator as $D_i=I(T_i\le C_i)$ for $i=1,2,\ldots,n$. Under the independence of the censoring and failure time variables, Kaplan and Meier [33] provided the estimator of the survival function given by

$$S(y)=\prod_{i:Y_i\le y}\left(\frac{n-R_i}{n-R_i+1}\right)^{d_i},$$

where R_i is the rank of the ith observation in the observed sample. This classic nonparametric estimator of the survival function is used as a standard for comparing the other estimated survival functions. Although the Kaplan-Meier estimator is ML for discrete distributions only, Johansen et al. [34] showed that it is a ML estimator in the class of all distributions under the generalized ML framework developed by Kiefer and Wolfowitz [35].

The R codes to obtain the K-M survival probabilities, to plot the K-M survival curves and to obtain the estimates of cure fraction from the K-M survival curves are provided in appendix.

4. Parametric maximum likelihood estimation

The likelihood function for an observed data, say $O = \{y_i, d_i, X_i, i = 1, 2, \ldots, n\}$, can be written as

$$l(\gamma, \beta|O) = \prod_{i=1}^{n} \left[f_p(y_i)\right]^{d_i} \left[S_p(y_i)\right]^{1-d_i} = \prod_{i=1}^{n} \left[h_p(y_i)\right]^{d_i} S_p(y_i),$$

where γ is a vector of baseline distribution parameters. Replacing $f_p(.)$ and $S_p(.)$ from (4) and (5), we have the likelihood function for PTCM based on generalized Gompertz distribution given by

$$l(\gamma, \beta|O) = \prod_{i=1}^{n} \exp\Big(- e^{X_i'\beta} \left(1 - \exp\Big(- \frac{\alpha}{\lambda}\left(e^{\lambda y_i} - 1\right)\Big)\right)^{\delta} \Big)$$

$$\times \left[\delta\alpha e^{X_i'\beta + \lambda y_i} \exp\Big(- \frac{\alpha}{\lambda}\left(e^{\lambda y_i} - 1\right)\Big) \left(1 - \exp\Big(- \frac{\alpha}{\lambda}\left(e^{\lambda y_i} - 1\right)\Big)\right)^{\delta - 1}\right]^{d_i}$$

$$= (\delta\alpha)^{\sum_{i=1}^{n} d_i} e^{\sum_{i=1}^{n} d_i \left(\sum_{j=0}^{k} B_j x_{ji}\right) + \lambda \sum_{i=1}^{n} d_i y_i} \exp\left(- \frac{\alpha}{\lambda} \sum_{i=1}^{n} d_i \left(e^{\lambda y_i} - 1\right)\right)$$

$$\times \exp\left(- \sum_{i=1}^{n} e^{\left(\sum_{j=0}^{k} B_j x_{ji}\right)} \left(1 - \exp\Big(- \frac{\alpha}{\lambda}\left(e^{\lambda y_i} - 1\right)\Big)\right)^{\delta}\right)$$

$$\times \prod_{i=1}^{n} \left(1 - \exp\Big(- \frac{\alpha}{\lambda}\left(e^{\lambda y_i} - 1\right)\Big)\right)^{(\delta - 1) d_i}.$$

The corresponding log-likelihood function, say $L(\gamma, \beta|O) = \ln[l(\gamma, \beta|O)]$, is

$$L(\gamma, \beta|O) = \ln(\delta\alpha) \sum_{i=1}^{n} d_i + \sum_{i=1}^{n} d_i \left(\sum_{j=0}^{k} B_j x_{ji}\right) + \lambda \sum_{i=1}^{n} d_i y_i - \frac{\alpha}{\lambda} \sum_{i=1}^{n} d_i \left(e^{\lambda y_i} - 1\right)$$

$$+ (\delta - 1) \sum_{i=1}^{n} d_i \ln\left(1 - \exp\Big(- \frac{\alpha}{\lambda}\left(e^{\lambda y_i} - 1\right)\Big)\right)$$

$$- \sum_{i=1}^{n} e^{\left(\sum_{j=0}^{k} B_j x_{ji}\right)} \left(1 - \exp\Big(- \frac{\alpha}{\lambda}\left(e^{\lambda y_i} - 1\right)\Big)\right)^{\delta},$$

where k denotes the number of covariates, x_0 the vector of 1's, x_j's the covariates, B_0 the intercept term and B_j's the regression coefficients.

The likelihood function corresponding to PTCM based on generalized Weibull distribution can be obtained from (8) and (9) as

$$l(\gamma, \beta|O) = \prod_{i=1}^{n} \left[h_p(y_i)\right]^{d_i} S_p(y_i) = \prod_{i=1}^{n} \left[\exp(X_i'\beta)\alpha\lambda\delta y_i^{\alpha - 1} \exp(-\lambda y_i^{\alpha}) \left(1 - \exp(-\lambda y_i^{\alpha})\right)^{\delta - 1}\right]^{d_i}$$

$$\times \exp\Big(- \exp(X'\beta) \left(1 - \exp(-\lambda y_i^{\alpha})\right)^{\delta} \Big)$$

$$= \prod_{i=1}^{n} e^{\left(\sum_{j=0}^{k} B_j x_j\right) d_i} (\alpha\lambda\delta)^{d_i} y_i^{(\alpha - 1) d_i} \exp(-\lambda d_i y_i^{\alpha}) \left(1 - \exp(-\lambda y_i^{\alpha})\right)^{(\delta - 1) d_i}$$

$$\times \exp\left(- e^{\sum_{j=0}^{k} B_j x_j} \left(1 - \exp(-\lambda y_i^{\alpha})\right)^{\delta}\right)$$

The corresponding log-likelihood function is

$$L(\gamma,\beta|O)=\ln(\alpha\lambda\delta)\sum_{i=1}^{n}d_i+\sum_{i=1}^{n}d_i\left(\sum_{j=0}^{k}B_jx_{ji}\right)+(\alpha-1)\sum_{i=1}^{n}d_i\ln y_i$$

$$-\lambda\sum_{i=1}^{n}d_iy_i^{\alpha}+(\delta-1)\sum_{i=1}^{n}d_i\ln\left(1-\exp\left(-\lambda y_i^{\alpha}\right)\right)$$

$$-\sum_{i=1}^{n}e^{\left(\sum_{j=0}^{k}B_jx_{ji}\right)}\left(1-\exp\left(-\lambda y_i^{\alpha}\right)\right)^{\delta}.$$

Similarly, the likelihood function corresponding to PTCM based on generalized gamma distribution can be obtained.

The maximum likelihood estimate $(\widehat{\gamma},\widehat{\beta})$ of (γ,β) can be obtained by maximizing the log-likelihood function $L(\gamma,\beta|O)$. It is well known that the derivation of expected Fisher information matrix is not possible, we can use the observed information matrix as an estimate of the expected Fisher information matrix. Under standard regularity conditions of ML estimators, we can state that the vector $(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta})\sim\boldsymbol{N}(\boldsymbol{0},\boldsymbol{V})$, where $\boldsymbol{\vartheta}=(\gamma,\beta)'$, $\widehat{\boldsymbol{\vartheta}}=(\widehat{\gamma},\widehat{\beta})'$ and $\boldsymbol{V}$ is the inverse of negative of the Hessian matrix evaluated at $\widehat{\boldsymbol{\vartheta}}$. The R codes to fit the promotion time cure model by the method of parametric ML estimation and to plot the corresponding estimated survival function are provided in appendix.

5. Semiparametric maximum likelihood estimation

For semiparametric estimation, the distribution function $F_Z(.)$ of the random variable Z is left unspecified and estimated non-parametrically. The nonparametric estimation of F, discussed in Ref. [6], requires the determination of a threshold such that all censored observations greater than this threshold are treated as $y_i=\infty$ and all other observations lower than the threshold are treated as $y_i<\infty$. This is necessary because one can choose $f(y_i)=\infty$ for some y_i with $d_i=1$. However, if a parametric form is assumed for $F_Z(.)$ as in Section 4, then this condition is not needed. Following [6,36], the likelihood function for an observed data $O_i=(y_i,d_i,X_i)$; $i=1,2,\ldots,n,$ can be written as

$$l(\beta,F|O)=\prod_{i=1}^{n}\left\{\left[\left(\exp(X_i'\beta)F\{y_i\}\right)^{d_i}\exp\left(-\exp(X_i'\beta)F(y_i)\right)\right]^{I(y_i<\infty)}\right.$$

$$\left.\times\left[\exp\left(-\exp(X_i'\beta)\right)\right]^{I(y_i=\infty)}\right\},$$

where $F\{y\}$ denotes the jump size of $F(.)$ at y and $F(.)$ is the right continuous distribution function with jumps at event times only. The corresponding log-likelihood function is

$$L(\beta,F|O)=\sum_{i=1}^{n}I(y_i<\infty)d_i\left[(X_i'\beta)+\ln(F\{y_i\})\right]-\sum_{i=1}^{n}I(y_i<\infty)\left[\exp(X_i'\beta)F(y_i)\right].$$

Denoting $Y_{(1)},\ldots,Y_{(m)}$ as the ordered distinct failure times and $p_{(1)},\ldots,p_{(m)}$ as the corresponding jump sizes such that $\sum_i^m p_{(i)}=1$, where m is the number of distinct failure times, the Lagrange multiplier constraint log-likelihood function can be written as

$$L(\beta,F|O)=\sum_{i=1}^{n}I(y_i<\infty)d_i\left[X_i'\beta+\ln(p_i)\right]$$

$$-\sum_{i=1}^{n}I(y_i<\infty)\left[\exp\left(X_i'\beta\right)F(y_i)\right]-n\lambda\left(\sum_{i}^{m}p_{(i)}-1\right),\tag{10}$$

where

$$p_{(i)}=F\{y_i\},F(y_i)=\sum\nolimits_{y_j\le y_i,d_j=1}p_j\ \text{ and }\ F(\infty)=1.$$

Differentiating (10) with respect to $p_{(i)}$, λ, β and equating to zero the resulting expressions, we have

$$\frac{\partial L}{\partial p_{(i)}}=\frac{1}{p_{(i)}}-\sum_{j=1}^{n}I\left(y_{(i)}\le y_j<\infty\right)\left[\exp\left(X_i'\beta\right)\right]-n\lambda=0,\tag{11}$$

$$\frac{\partial L}{\partial \lambda} = n\left(\sum_{i}^{m} p_{(i)} - 1\right) = 0, \tag{12}$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} I(y_i < \infty) d_i X_i - \sum_{i=1}^{n} \left[\exp\left(X_i'\beta\right) F(y_i)\right] X_i = 0. \tag{13}$$

The maximum likelihood estimates $(\widehat{\lambda}, \widehat{\beta})$ of (λ, β) can be obtained by first solving (11) and (12) for $p_{(i)}$'s and λ by fixing β; then solving (13) for β by fixing $p_{(i)}$'s and λ using some nonlinear iterative procedure. The R codes to fit the promotion time cure model by the method of semiparametric ML estimation and to plot the corresponding estimated survival function are provided in appendix.

6. Real data analysis

Gastric cancer is one of the most common cancer diseases worldwide and despite a decline in death due to stomach cancer during the last few decades, it is still one of the leading causes of cancer related death. According to the GLOBOCAN 2020 estimates, stomach cancer caused approximately 800000 deaths (accounting for 7.7 % of all cancer deaths) and ranks as the fourth leading cause of cancer deaths in both genders combined. About 1.1 million new cases of stomach cancer were diagnosed in 2020 (accounting for 5.6 % of all cancer cases). About 75 % of all new cases and all deaths from stomach cancer are reported in Asia. These facts highlight the need for cancer research worldwide.

The real data used in this study belong to a retrospective study in patients with gastric adenocarcinoma conducted by Jácome et al. [37] between January 2002 and December 2007. There are 76 patients treated with adjuvant chemoradiotherapy (CRT) and 125 patients treated with surgery alone. About 58 % of the data are censored in CRT group and about 50 % in the surgery alone group. The same data was analyzed by Martinez et al. [38] in Bayesian context. First, we draw Kaplan–Meier survival curves in Fig. 1 and apply log-rank test of no difference of survival experiences between the two groups.

of patients. We can see that the survival curves level off at a time substantially greater than 0 after 26- or 30- month-follow-up at survival probabilities significantly greater from zero for both the treatment groups. This implies that some of the patients in both groups were cured and did not experience the event following treatments. The Kaplan–Meier estimates of cure fraction obtained from the Kaplan–Meier survival curves are 0.473 for surgery alone group and 0.545 for CRT group. The p-value of log-rank test is 0.04 which implies that the two groups of patients face different survival experiences. Next, we consider the semiparametric ML method to fit the PTCM using a binary covariate X taking a value 1 for CRT treatment group and 0 for the surgery alone treatment group. The estimate of the cure fraction is 0.4307 for surgery alone group which is smaller than the estimate given by Kaplan–Meier survival curve and 0.5585 for CRT group which is larger than the estimate given by Kaplan–Meier survival curve. The population survival function estimated by semiparametric ML method is plotted in Fig. 1 which shows a poor fit to Kaplan–Meier survival curves. In general, the semiparametric estimate of survival function fits well to Kaplan–Meier survival curve, however, here the performance of semiparametric ML method is very poor.

Now we compute the ML estimates, standard errors (SEs) and the corresponding AIC for the PTCM based on standard distributions and report the results in Table 1. We can see that the results for cure fractions p_0 and p_1 are very poor in case of the PTCM based on Weibull and gamma distributions and are very good in case of the PTCM based on Gompertz distribution regarding K-M nonparametric criterion. That is, the cure fraction estimates are relatively close to the estimates given by Kaplan–Meier method for standard Gompertz cure model and far from the estimates given by Kaplan-Meier method for standard Weibull and gamma cure models. Also, in terms of

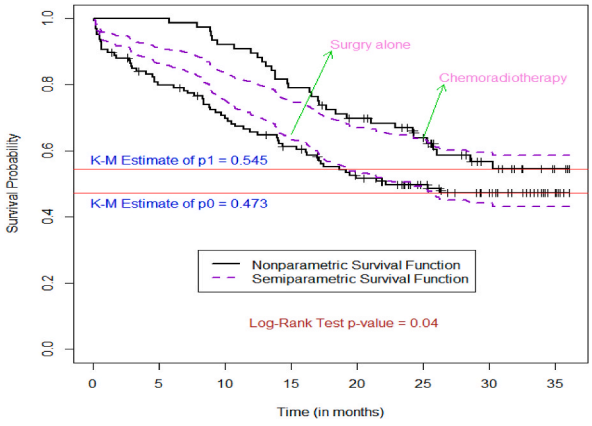


Fig. 1. Plot of the estimated nonparametric and semiparametric survival functions and the Kaplan–Meier survival estimates of the cure fractions.

Table 1
The maximum likelihood estimates of model parameters, cure fractions and the associated goodness of fit measure for standard cure models.

Model	Parameter	Estimate	SE	
Gompertz	b_0	−0.1636	0.1359	AIC = 893.80 $p_0 = 0.4278$
	b_1	−0.4613	0.2175	$p_1 = 0.579$
	α	0.0298	0.0064	
	λ	0.0712	0.0193	
Weibull	b_0	1.3509	2.3777	AIC = 898.16 $p_0 = 0.0211$
	b_1	−0.4474	0.2180	$p_1 = 0.0847$
	α	0.8747	0.1524	
	λ	0.0116	0.0269	
Gamma	b_0	1.0182	1.6014	AIC = 898.1 $p_0 = 0.0627$
	b_1	−0.4470	0.2179	$p_1 = 0.1702$
	α	0.8778	0.1409	
	λ	0.0086	0.0200	
Semiparametric	b_0	−0.1715	0.1357	$p_0 = 0.4307$
	b_1	−0.4523	0.2076	$p_1 = 0.55851$

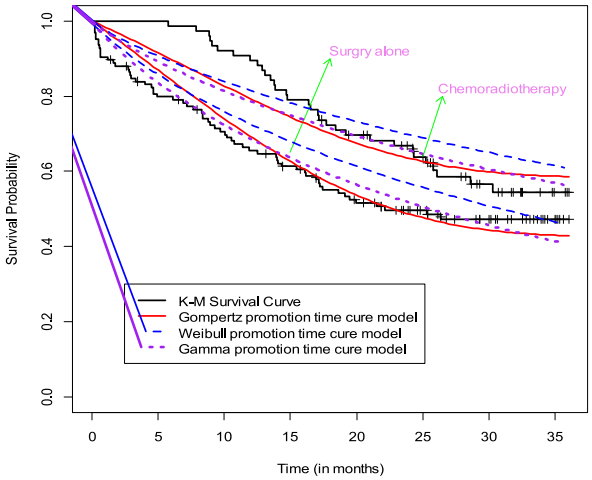


Fig. 2. The plot of the estimated survival functions for the standard cure models fitted to K-M survival curves.

Table 2
The maximum likelihood estimates of model parameters, cure fractions and the associated goodness of fit measure for generalized cure models.

Model	Parameter	Estimate	SE	
Generalized	b_0	−0.1794	0.1326	AIC = 888.72 $p_0 = 0.4335$
Gompertz	b_1	−0.4538	0.2180	$p_1 = 0.5881$
	δ	0.6167	0.1170	
	α	0.0094	0.0057	
	λ	0.1304	0.0308	
Generalized	b_0	−0.1733	0.1431	AIC = 895.54 $p_0 = 0.4313$
Weibull	b_1	−0.4468	0.2180	$p_1 = 0.5840$
	δ	0.2602	0.0281	
	α	3.6010	0.0047	
	λ	7.05e-6	3.79e-6	
Generalized	b_0	−0.1911	0.1324	AIC = 892.5 $p_0 = 0.4378$
Gamma	b_1	−0.4685	0.2179	$p_1 = 0.5901$
	δ	0.0535	0.0636	
	α	16.5821	19.128	
	λ	29.9170	1.5731	

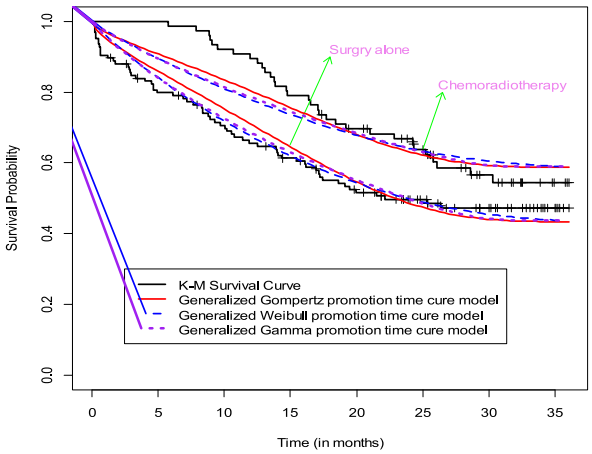


Fig. 3. The plot of the estimated survival functions for the generalized cure models fitted to K-M survival curves.

Table 3
The maximum likelihood estimates of model parameters, cure fractions and the associated goodness of fit measure for generalized cure models.

Model	Parameter	Estimate	SE	
Generalized	b_0	−0.2870	0.1326	AIC = 868.99 $p_0 = 0.4272$
Gompertz	b_1	−0.2175	0.2245	$p_1 = 0.5467$
	B_0	−0.4062	0.1916	
	B_1	1.1423	0.2398	
	α	0.0266	0.0131	
	λ	0.0882	0.0273	
Generalized	b_0	−0.2993	0.1325	AIC = 869.71 $p_0 = 0.4765$
Weibull	b_1	−0.1861	0.2243	$p_1 = 0.5404$
	B_0	−1.3995	0.1213	
	B_1	1.2475	0.2143	
	α	3.2223	0.0427	
	λ	4.58e-5	1.37e-5	
Generalized	b_0	−0.2992	0.1321	AIC = 869.40 $p_0 = 0.4764$
Gamma	b_1	−0.1978	0.2249	$p_1 = 0.5442$
	B_0	−1.9443	0.0796	
	B_1	1.1466	0.2054	
	α	5.2768	2.6149	
	λ	26.1774	2.6131	

AIC, the performance of PTCM based on Gompertz distribution is far better than the PTCM based on Weibull and gamma distributions. Based on these results, we can say that the PTCM based on standard Gomperz distribution can provide better fit to the data with fish-shaped survival functions than semiparametric promotion time cure model and the parametric promotion time cure model based on standard Weibull and gamma distributions. However, neither of the model fit the data reasonably well as indicated by the survival functions plotted in Fig. 2. Next, we compute the ML estimates, standard errors, and the corresponding AIC for the PTCM based on generalized distributions and provide the results in Table 2. We see

that the results for cure fractions p_0 and p_1 are improved and approximately the same for all the three models and their goodness of fit have improved reasonably. Based on these results, we can say that the PTCM based on generalized distributions can provide better fit to the data than the parametric PTCM based on standard distributions. Still. however, neither of the models fit the data reasonably well as indicated by the survival functions plotted in Fig. 3. The reason may be that the arms of the data have very different hazard rates, one has monotone decreasing hazard rate and the other has increasing-decreasing hazard rate. To overcome the structure of different curve shapes of hazard rates between the two arms of data, we introduce the covariate in shape parameter delta for all the three generalized distributions. The ML estimates, standard errors and the corresponding AIC for these models are given in Table 3. We see that the AIC goodness of fit measure for the three models have largely further improved with lowest AIC for the generalized Gompertz cure model. To visualize the fitness of the models to real data, we plot the estimated PTCM survival function with different baseline distributions in Fig. 4. We see that now the estimated population survival function of the PTCM based on all the three distributions fit the Kaplan–Meier survival curves equally well with slightly better fit for the generalized Gompertz cure model. Based on

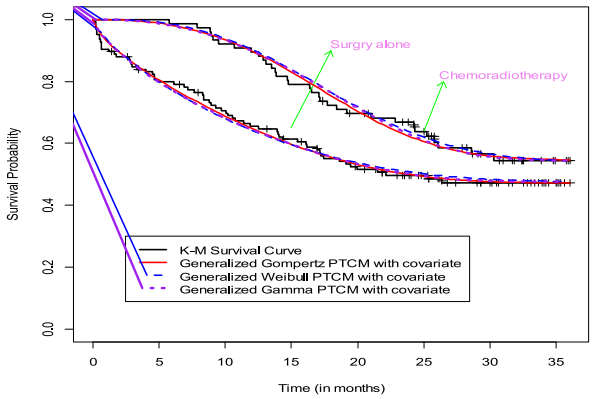


Fig. 4. The plot of estimated survival functions for the generalized cure models fitted to K-M survival curves when the covariate is introduced to the shape parameter delta.

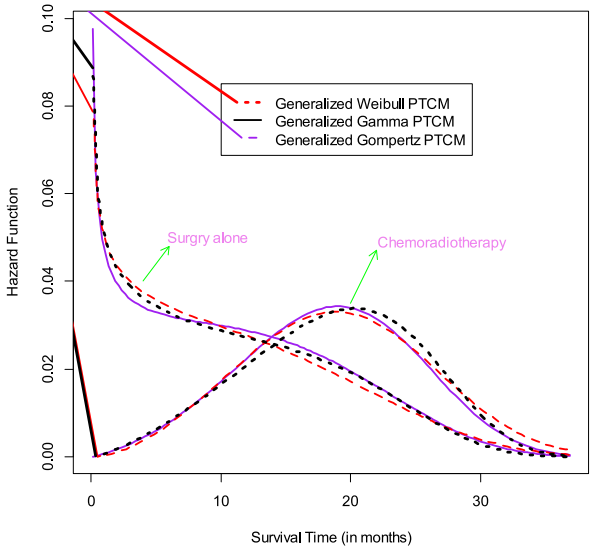


Fig. 5. The estimated hazard functions of the promotion time cure models based on different distributions for the surgery alone and chemoradiotherapy groups.

the forgoing results, we can say that the generalized Gompertz cure model can be used as an alternative to the generalized Weibull and generalized gamma cure models to analyze the survival data with cure fraction. Before drawing the conclusions, we plot the estimated population hazard functions corresponding to three generalized PTCMs in Fig. 5.

7. Conclusions

Several distributions including exponential, gamma, Weibull, lognormal, log-logistic, Burr type XII, etc., are used as baseline distributions for cure models. In this article, we study promotion time cure model with generalized Gompertz as the baseline distribution. A simulation study is performed to compare the goodness of fit of the proposed model with generalized Weibull and gamma promotion time cure models. It is observed that there is not much difference among the results for the three different models based on AIC. The real data used in this study belong to a retrospective study in patients with gastric adenocarcinoma. There are 76 patients treated with adjuvant chemoradiotherapy (CRT) and 125 patients treated with surgery alone. The p-value of log-rank test is 0.04 which implies that the two groups of patients face different survival experiences. It is seen that the semiparametric estimate of survival function fits poorly to the Kaplan–Meier survival curve. In terms of AIC, the performance of the promotion time cure model based on Gompertz distribution is far better than the promotion time cure model based on Weibull and gamma distributions. When we fit the generalized promotion time cure models, we see that the results for cure fractions p_0 and p_1 are improved and these are approximately the same for all the three models and their goodness of fit have improved reasonably. It is observed that the promotion time cure model based on generalized distributions provides better fit to the data than the parametric promotion time cure model

based on standard distributions. However, neither of the model fit the data reasonably well as indicated by the survival functions

fitted to K-M survival curve. The reason may be that the arms of the data have very different hazard rates, one has monotone decreasing hazard rate and the other has increasing-decreasing hazard rate. To overcome the structure of different curve shapes of hazard rates between the two arms of data, we introduce the covariate in shape parameter delta for all the three generalized distributions. We see that now the estimated population survival function of the promotion time cure model based on all the three distributions fit the Kaplan–Meier survival curves equally well with slightly better fit for the generalized Gompertz cure model. It is concluded that the generalized Gompertz cure model can be used as an alternative to the generalized Weibull and generalized gamma cure models to analyze the survival data with cure fraction. It is further inferred from the data at hand that the estimated survival function corresponding to CRT treatment is higher than the estimated survival function corresponding to surgery alone treatment during the study period. The estimated hazard function corresponding to the patients treated with CRT is under the estimated hazard function corresponding to the patients treated with surgery alone during the follow-up time 0 month to 14 months, and above after this interval. The log rank test of no difference of survival experience between two treatment groups of patients is statistically significance with p-value 0.04. We carried out a limited simulation study to observe the goodness of fit of the proposed model in comparison to the other generalized promotion time cure model models, the results are very encouraging. However, an extensive simulation study is required to observe the behavior of the estimators of its model parameters and to compare with other generalized cure models.

Funding declaration

The authors declare that this research did not receive any research grant.

Data availability

The data used to support research results belong to a retrospective study in patients with gastric adenocarcinoma conducted by Jácome et al. [37] between January 2002 and December 2007 and is available in Martinez et al. [38].

Additional information

No additional information is available about this paper.

Ethical approval and consent to participate

Not applicable.

CRediT authorship contribution statement

Ayesha Tahira: Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Muhammad Yameen Danish:** Supervision, Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

R codes for fitting parametric promotion time cure model based on the generalized distributions and semiparametric promotion time cure model.

R codes for fitting parametric promotion time cure model based on the generalized distributions and semiparametric promotion time cure model

(a) Generalized Gompertz

```
library(maxLik)
   LLF <- function(pars, data) {
 paras <- pars[1:6]
    b0 <- pars[1]
    b1 <- pars[2]
    B0 <- pars[3]
    B1 <- pars[4]
 alpha <- pars[5]
  beta <- pars[6]
     p <- exp(-exp(b0+b1*z))
 delta <- exp(B0+B1*z)
    F1 <- 1-exp(-(alpha/beta)*exp(beta*x)+(alpha/beta))
     F <- F1^delta
     f <- alpha*delta*exp(beta*x)*(1-F1)*F1^(delta-1)
     L <- (-log(p))^d*f^d*p^F
   LLF <- sum(log(L))
        }
GGomfit <- maxLik(LLF, start = c(b0, b1, B0, B1, alpha,
                  beta), data = data)
GGomfit
```

(b) Generalized Weibull

```
   library(maxLik)
   LLF <- function(pars, data) {
 paras <- pars[1:6]
    b0 <- pars[1]
    b1 <- pars[2]
    B0 <- pars[3]
    B1 <- pars[4]
 alpha <- pars[5]
lambda <- pars[6]
    p <- exp(-exp(b0+b1*z))
delta <- exp(B0+B1*z)
   F1 <- 1-exp(-lambda*x^alpha)
    F <- F1^delta
    f <- alpha*delta*lambda*x^(alpha-1)*(1-F1)*F1^(delta-1)
    L <- (-log(p))^d*f^d*p^F
    LLF <- sum(log(L))
                  }
```

```
GWeifit <- maxLik(LLF, start = c (b0, b1, B0, B1, alpha,
           lambda), data = data)
GWeifit
```

(c) Generalized Gamma

```
library(flexsurvcure)
GGamfit <- flexsurvcure(Surv(x, d)~z, data = data, link =
         "loglog", dist = "gengamma.orig", mixture=F)
GGamfit
```

(d) Semiparametric based on backward fitting

```
library(miCoPTCM)
vc <- matrix(nrow = 2,ncol = 2,0)
Semptcm <- PTCMestimBF(formula = Surv(x, d)~z, data = data,
     varCov = vc,init = runif(2))
summary(Semptcm)
SDF <- Semptcm$estimCDF
```

(e) R code for estimating and plotting nonparametric K-M and semiparametric survival functions

```
library(survival)
KM <- survfit(Surv(x, d)~z, data = data)
plot(KM, col = c("black", "Black"), lwd = rep(2, 2), lty =
     c(1, 1), mark.time = TRUE, xlab = "Time (in months)",
     ylab = "Survival Probability")
legend(list(x = 8,y = 0.30), legend = c("Nonparametric
     Survival Function", "Semiparametric Survival
     Function"), col = c("black", "purple"),
     lwd = c(2, 2), lty = c(1, 2))
text(19, 0.08, col = "brown", expression("Log-Rank Test
     p-value = 0.04"))
abline(h = 0.545, col = "red", lty = 1, lwd = 1)
abline(h = 0.473, col = "red", lty = 1, lwd = 1)
text(6.5, 0.57, col="blue", expression("K-M Estimate of
     p1 = 0.545"))
text(6.5, 0.44, col="blue", expression("K-M Estimate of
     p0 = 0.473"))
arrows(x0 = 15, y0 = 0.65, x1 = 18, y1 = 0.90,
     lwd = 1, col="green", length=0.1)
arrows(x0 = 25, y0 = 0.64, x1 = 26.5, y1 = 0.80, lwd = 1,
     col = "green", length=0.1)
text(21, 0.92, col = "violet", expression("Surgry alone"))
text(31, 0.82, col = "violet",
     expression("Chemoradiotherapy"))
```

. (*continued*).

```
SF0 <- exp(-exp(b0*SDF)
SF1 <- exp(-exp(b0+b1)*SDF)
sx <- sort(x)
lines(sx, SF0, col = "purple", lwd = 2, lty = 2)
lines(sx, SF1, col = "purple", lwd = 2, lty = 2)
```

. (*continued*).

References

[1] J.B. Boag, Maximum likelihood estimates of the proportion of patients cured by cancer therapy, J. Royal Stat. Soc. Ser. B 11 (1949) 15–44. https://www.jstor.org/stable/2983694.

[2] J. Berkson, R.P. Gage, Survival curve for cancer patients following treatment, J. Am. Stat. Assoc. 47 (1952) 501–515. https://www.jstor.org/stable/2281318.

[3] V.T. Farewell, The use of mixture models for the analysis of survival data with long-term survivors, Biometrics 38 (1982) 1041–1046. https://www.jstor.org/stable/2529885.

[4] A.Y. Yakovlev, A.D. Tsodikov, B. Asselain, Stochastic Models of Tumor Latency and Their Biostatistical Applications, Vol. 1 of Mathematical Biology and Medicine, World Scientfic, Singapore, 1996, https://doi.org/10.1142/2420.

[5] M.H. Chen, J.G. Ibrahim, D. Sinha, A new Bayesian model for survival data with a survival fraction, J. Royal Stat. Soc. Ser. C 94 (1999) 909–919, https://doi.org/10.1080/01621459.1999.10474196.

[6] D. Zeng, G. Yin, J.G. Ibrahim, Semiparametric transformation models for survival data with a cure fraction, J. Am. Stat. Assoc. 101 (2006) 670–684, https://doi.org/10.1198/016214505000001122.

[7] P.C. Lambert, J.R. Thompson, C.L. Weston, P.W. Dickman, Estimating and modeling the cure fraction in population-based cancer survival analysis, Biostatistics 8 (2006) 576–594, https://doi.org/10.1093/biostatistics/kxl030.

[8] G. Yin, J.G. Ibrahim, Cure rate model: a unified approach, Can. J. Statist. 33 (2005) 559–570. https://www.jstor.org/stable/25046202.

[9] J.M.G. Taylor, N. Liu, Statistical issues involved with extending standards models, in: V. Nair (Ed.), Advances in Statistical Modelling and Inference: Essays in Honor of Kjell A. Doksum, Series in Biostatistics, World Scientfic, Singapore, 2007, pp. 299–311, https://doi.org/10.1142/9789812708298_0015 (Chapter 15).

[10] Y. Peng, J. Xu, An extended cure model and model selection, Lifetime Data Anal. 18 (2012) 215–233, https://doi.org/10.1007/s10985-011-9213-1.

[11] M.H. Chen, J.G. Ibrahim, Maximum likelihood methods for cure rate models with missing covariates, Biometrics 57 (2001) 43–52, https://doi.org/10.1111/j.0006-341x.2001.00043.x.

[12] A. D'Andrea, R. Rocha, V. Tomazella, F. Louzada, Negative binomial Kumaraswamy-G cure rate regression model, J. Risk. Financial Manag. 11 (2018) 1–14, https://doi.org/10.3390/jrfm11010006.

[13] A.D. Tsodikov, J.G. Ibrahim, A.Y. Yakovlev, Estimating cure rates from survival data, J. Am. Stat. Assoc. 98 (2003) 1063–1078, https://doi.org/10.1198/016214503000000001007.

[14] M. Castro, V.G. Cancho, J. Rodrigues, Bayesian long-term survival model parametrized in the cured fraction, Biom. J. 51 (2009) 443–455, https://doi.org/10.1002/bimj.200800199.

[15] H. Seltman, J. Greenhouse, L. Wasserman, Bayesian model selection: analysis of a survival model with a surviving fraction, Stat. Med. 20 (2001) 1681–1691, https://doi.org/10.1002/sim.779.

[16] N. Balakrishnan, S. Pal, *Likelihood inference for flexible cure* rate models with gamm*a lifetimes*, Comm. Statist. Theory Methods 19 (2015) 4007–4048, https://doi.org/10.1080/03610926.2014.964807.

[17] J. Leao, M. Bourguignon, H. Saulo, M. Santos-Neto, V. Calsavara, *T*he n*egative binomial beta prime regression model with cure rate: application w*ith a m*elanoma dataset*, J. Stat. Theory. Pract. 15 (2021) 1–21, https://doi.org/10.1007/s42519-021-00195-y.

[18] K. Abbas, N.Y. Abbasi, A. Ali, S.A. Khan, S. Manzoor, A. Khalil, U. Khalil, D.M. Khan, Z. Hussain, M. Altaf, Bayesian analysis of three-parameter Fréchet distribution with medical applications, Comput. Math. Methods Med. (2019), https://doi.org/10.1155/2019/9089856.

[19] Y. Liu, M. Ilyas, S.K. Khosa, E. Muhmoudi, Z. Ahmad, D.M. Khan, G.G. Hamedani, A flexible reduced logarithmic-X family of distributions with biomedical analysis, Comput. Math. Methods Med. (2020), https://doi.org/10.1155/2020/4373595.

[20] K. Abbas, Z. Hussain, N. Rashid, A. Ali, M. Taj, S.A. Khan, S. Manzoor, U. Khalil, D.M. Khan, Bayesian estimation of Gumbel type-II distribution under type-II censoring with medical applications, Comput. Math. Methods Med. (2020) 1–11, https://doi.org/10.1155/2020/1876073.

[21] M. Pedrosa-Laza, A. López-Cheda, R. Cao, Cure models to estimate time until hospitalization due to COVID-19. A case study in Galicia, Appl. Intell. 52 (2022) (2020) 794–807, https://doi.org/10.1007/s10489-021-02311-8.

[22] L. Botta, J. Goungounga, R. Capocaccia, G. Romain, M. Colonna, G. Gatta, O. Boussari, V. Jooste, A new cure model that corrects for increased risk of non-cancer death: analysis of reliability and robustness, and application to real-life data, BMC Med. Res. Methodol. 23 (2023) 1–19, https://doi.org/10.1186/s12874-023-01876-x.

[23] M. Escobar-Bach, I. Van Keilegom, Nonparametric estimation of conditional cure models for heavy-tailed distributions and under insufficient follow-up, Comput. Statist. Data Anal. 183 (2023) 107728, https://doi.org/10.1016/j.csda.2023.107728.

[24] A. Ezquerro, B. Cancela, A. López-Cheda, On the reliability of machine learning models for survival analysis when cure is a possibility, Mathematics 11 (2023) 4150, https://doi.org/10.3390/math11194150.

[25] A. Ei-Gohary, A. Alshamrani, A.N. Al-Otaibi, The generalized Gompertz distribution, Appl. Math. Model. 37 (2013) 13–24, https://doi.org/10.1016/j.apm.2011.05.017.

[26] M. Obeidat, A. Al-Nasser, A.I. Al-Omari, Estimation of generalized Gompertz distribution parameters under ranked-set sampling, j. probab. stat. 2020 (7362657) (2020) 1–14, https://doi.org/10.1155/2020/7362657.

[27] A. Martinez, Defective generalized Gompertz distribution and its use in the analysis of lifetime data in presence of cure fraction, censored data and covariates, Electron. J. Appl. Stat. Anal. 10 (2017) 463–484. http://creativecommons.org/licenses/by-nc-nd/3.0/it/.

[28] P. Borges, 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, J. Stat. Comput. Simul. 87 (2017) 1712–1722, https://doi.org/10.1080/00949655.2017.1281927.

[29] E. Demir, B. Saracoglu, Maximum likelihood estimation for the parameters of the generalized Gompertz distribution under progressive type-ii right censored samples, j. appl. nat. sci. 4 (2015) 41–48. https://www.researchgate.net/publication/274068697.

[30] M.M. Nassar, F.H. Eissa, On the exponentiated Weibull distribution, Commun. Stat.-Theory Methods 32 (2003) 1317–1336, https://doi.org/10.1081/STA-120021561.

[31] E.W. Stacy, A generalization of gamma distribution. Annals of Mathematical Statistics 33 (1962) 1187–1192, https://doi.org/10.1214/aoms/1177704481.

[32] J.F. Lawless, Inference in the generalized gamma and log gamma distributions, Technometrics 22 (1980) 409–419, https://doi.org/10.2307/1268326.

[33] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Am. Stat. Assoc. 53 (1958) 457–481, https://doi.org/10.2307/2281868.

[34] S. Johansen, The product limit estimator as maximum likelihood estimator, Scand. J. Stat. 5 (1978) 195–199. https://www.jstor.org/stable/4615715.

[35] J. Kiefer, J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Ann. Math. Stat. 27 (1956) 887–906. https://www.jstor.org/stable/2237188.

[36] Y. Ma, G. Yin, Cure rate model with mismeasured covariates under transformation, J. Am. Stat. Assoc. 103 (2008) 743–756, https://doi.org/10.1198/016214508000000319.

[37] A.A. Jácome, D.R. Wohnrath, C. Scapulatempo Neto, J.H. Fregnani, A.L. Quinto, A.T. Oliveira, V.L. Vazquez, G. Fava, E.Z. Martinez, J.S. Santos, Effect of adjuvant chemoradiotherapy on overall survival of gastric cancer patients submitted to D2 lymphadenectomy, Gastric Cancer 16 (2013) 233–238, https://doi.org/10.1007/s10120-012-0171-4.

[38] E.Z. Martinez, J.A. Achcar, A.A.A. Jácome, J.S. Santos, Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data, Comput. Methods Programs Biomed. 1 (1 2) (2013) 343–355, https://doi.org/10.1016/j.cmpb.2013.07.021.