New defective models based on the Kumaraswamy family of distributions with application to cancer data sets

Ricardo Rocha,¹ Saralees Nadarajah,²
Vera Tomazella,¹ Francisco Louzada³ and Amanda Eudes¹

Statistical Methods in Medical Research
0(0) 1–23

© The Author(s) 2015

Reprints and permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/0962280215587976

smm.sagepub.com

Abstract

An alternative to the standard mixture model is proposed for modeling data containing cured elements or a cure fraction. This approach is based on the use of defective distributions to estimate the cure fraction as a function of the estimated parameters. In the literature there are just two of these distributions: the Gompertz and the inverse Gaussian. Here, we propose two new defective distributions: the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions, extensions of the Gompertz and inverse Gaussian distributions under the Kumaraswamy family of distributions. We show in fact that if a distribution is defective, then its extension under the Kumaraswamy family is defective too. We consider maximum likelihood estimation of the extensions and check its finite sample performance. We use three real cancer data sets to show that the new defective distributions offer better fits than baseline distributions.

Keywords

Cure fraction, defective distributions, Gompertz distribution, inverse Gaussian distribution, Kumaraswamy family, survival analysis

1 Introduction

In survival data, the goal is to model the time until the occurrence of an event of interest. In the general theory, it is assumed that an element is always susceptible to a certain event of interest that will eventually occur. However, such an assumption is often invalid because many databases have what we call immune elements. For immune elements, the event will never occur because they are not susceptible to the event of interest. Our interest in this paper is to model the fraction of immune elements, sometimes referred to as the cure fraction or the fraction of long-term survivors.

¹Universidade Federal de São Carlos, Departamento de Estatística, São Carlos, SP, Brasil

²University of Manchester, School of Mathematics, Manchester, UK

³Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil

Corresponding author:

Ricardo Rocha, Universidade Federal de São Carlos, Departamento de Estatística, São Carlos, SP, Brasil.

Email: ricardorocha23@hotmail.com

The usual and the most popular way to solve such problems is by a standard mixture model, first proposed by Boag¹ and Berkson and Gage.² The survival function is adjusted to $S(t) = p + (1 - p)S_0(t)$ such that $S_0(t)$ is a proper survival function. In this way, $S(t)$ converges to p as the time goes by. Berkson and Gage² analyzed a stomach cancer data set using the standard mixture model. Since then many other data sets have been analyzed using the standard mixture model. The most common choices for $S_0(t)$ have been the Weibull, log-logistic and lognormal distributions. Recently, different distributions have been proposed for this purpose.^{3–5}

Another way is to use defective distributions. These distributions are obtained from standard distributions by changing the domain of the parameters of the latter in such a way that their survival functions limit to $p \in (0, 1)$. This definition appears in the work of Balka et al.,^{6,7} but there are previous studies that use this idea. There are at least two distributions that can be modified to being defective: the Gompertz distribution⁸ and the inverse Gaussian distribution.⁷

This paper uses the Kumaraswamy family of distributions⁹ as a means for developing distributions that can be defective. Given a baseline distribution, this family adds two parameters, making it more flexible.

The Gompertz distribution is one that can take a defective form. For negative values of the shape parameter, the distribution becomes defective. Haybittle¹⁰ fitted the Gompertz distribution to a breast cancer data set. Cantor and Shuster¹¹ fitted a modified version of the same distribution to a pediatric cancer data set. Gieser et al.⁸ extended the distribution to include covariates.

The inverse Gaussian distribution is another distribution which can take a defective form. The inverse Gaussian distribution was derived by Schrödinger¹² by calculating the probability of the first time passage of a one-dimensional Brownian process (also called a Wiener process). A deeper study was made by Tweedie.¹³ It is also where the name inverse Gaussian arose.⁶ Defective versions of the inverse Gaussian distribution appear in Balka et al.^{6,7}

The main contributions of this paper are: (i) to show that if a distribution is defective, then its extension under the Kumaraswamy family of distributions is also defective; (ii) to propose two new defective distributions; (iii) to illustrate applications of these distributions using real cancer data sets with cure fraction.

The contents of this paper are organized as follows: the Gompertz and inverse Gaussian distributions, their probability density functions, their survival functions, their hazard functions, their defective versions and their cure fractions are presented in Section 2. Also presented in Section 2 are: (i) the Kumaraswamy family of distributions; (ii) two distributions, the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions, based on the family; (iii) maximum likelihood estimation of the family; (iv) regression models based on the family. Section 3 assesses the finite sample performance and checks the asymptotes of the maximum likelihood estimators. Section 4 illustrates the usefulness of these two distributions as defective models using three real cancer data sets. The fit of these two distributions is compared to those of the Gompertz and inverse Gaussian distributions, the baseline distributions, and the standard mixture model. Section 5 contains concluding remarks.

2 Methodology

2.1 Gompertz and inverse Gaussian defective distributions

Defective distributions are characterized by having probability density functions that integrate to values less than 1.⁶ Such distributions have the potential to model data with a cure fraction. Two known distributions that can be defective are the Gompertz and inverse Gaussian distributions.

The Gompertz distribution is often used for modeling survival data in various areas of knowledge,⁸ especially where there is suspicion of an approximately exponential hazard. Its probability density function is

$$g(t) = be^{at} e^{-\frac{b}{a}(e^{at}-1)} \quad (1)$$

for $a > 0$, $b > 0$ and $t > 0$. Here, a is the scale parameter and b is the rate parameter. The corresponding survival function is

$$S(t) = P(T > t) = e^{-\frac{b}{a}(e^{at}-1)} \quad (2)$$

The corresponding hazard function is

$$h(t) = be^{at}$$

The defective Gompertz distribution is the Gompertz distribution that allows for negative values of a . In this scenario, the hazard function changes from an increasing shape to a decreasing shape. The cure fraction in the population is calculated as the limit the survival function:

$$\lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} e^{-\frac{b}{a}(e^{at}-1)} = e^{\frac{b}{a}} = p \in (0, 1) \quad (3)$$

The inverse Gaussian distribution has the probability density function

$$g(t) = \frac{1}{\sqrt{2b\pi t^3}} \exp\left\{-\frac{1}{2bt}(1-at)^2\right\} \quad (4)$$

for $a > 0$, $b > 0$ and $t > 0$. The corresponding survival function is

$$S(t) = 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b} \Phi\left(\frac{-1-at}{\sqrt{bt}}\right) \right] \quad (5)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable. The corresponding hazard function is

$$h(t) = \frac{g(t)}{S(t)} = \frac{\frac{1}{\sqrt{2b\pi t^3}} \exp\left\{-\frac{1}{2bt}(1-at)^2\right\}}{1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b} \Phi\left(\frac{-1-at}{\sqrt{bt}}\right) \right]}$$

The defective inverse Gaussian distribution is the inverse Gaussian distribution that allows for negative values of a . The corresponding cure fraction is:

$$\lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b} \Phi\left(\frac{-1-at}{\sqrt{bt}}\right) \right] = 1 - e^{2a/b} \quad (6)$$

If the estimate of a is negative then estimates of the cure fraction can be obtained from (3) and (6). Figures 1 and 2 illustrate possible shapes of the probability density, survival and hazard functions of the defective Gompertz and defective inverse Gaussian distributions.

Figure 1. Probability density, survival and hazard functions of the defective Gompertz distribution.

The great advantage of these distributions is that the proportion of cured is always estimated using a distribution with one parameter less than the standard mixture model, which brings plenty of benefits in terms of estimation. But these distributions may not be as flexible as the standard mixture model.

We have been able to find only these two distributions that can be adapted to being defective. This does not mean there are not others. In order to construct other distributions of this kind, we propose the use of the Kumaraswamy family of distributions.⁹

Figure 2. Probability density, survival and hazard functions of the defective inverse Gaussian distribution.

2.2 The Kumaraswamy family of distributions

The Kumaraswamy distribution due to Kumaraswamy¹⁴ has the probability density and cumulative distribution functions specified by

$$f(x) = ru x^{r-1} (1 - x^r)^{u-1}$$

$$F(x) = 1 - (1 - x^r)^u$$

for $r > 0$, $u > 0$ and $0 < x < 1$. Here, both r and u are shape parameters. This distribution is closely related to the beta distribution, but it is simpler. Its hazard function can be unimodal, uniantimodal,

increasing, decreasing and constant. This shows that the Kumaraswamy distribution can model a wide variety of data sets.

Cordeiro and de Castro⁹ proposed the Kumaraswamy family of distributions: given a baseline cumulative distribution function $G(x)$ with $g(x) = dG(x)/dx$ and $S(x) = 1 - G(x)$, they define the Kumaraswamy- G distribution as the one having the probability density, cumulative distribution, survival and hazard rate functions specified by

$$g^*(x) = urg(x)G(x)^{r-1}[1 - G(x)^r]^{u-1} \quad (7)$$

$$G^*(x) = 1 - [1 - G(x)^r]^u \quad (8)$$

$$S^*(x) = [1 - G(x)^r]^u = \{1 - [1 - S(x)]^r\}^u \quad (9)$$

$$h^*(x) = g^*(x)/S^*(x) \quad (10)$$

For every given G , this defines a family of distributions. Clearly, the Kumaraswamy- G distribution for $r = u = 1$ is the baseline distribution.

Several Kumaraswamy- G distributions have been proposed in the literature: the Kumaraswamy Gumbel distribution,¹⁵ the Kumaraswamy Weibull distribution,¹⁶ the Kumaraswamy Birnbaum-Saunders distribution,¹⁷ the Kumaraswamy log-logistic distribution,¹⁸ the Kumaraswamy generalized gamma distribution,¹⁹ the Kumaraswamy Burr XII distribution,²⁰ and many others.

We now state and prove the result that if G is defective, then G^* is also defective.

Theorem 2.1: *If $S(t)$ is a survival function of a defective distribution, then $S^*(t)$ is also a survival function of a defective distribution.*

Proof: Suppose the limit of $S(t)$ is equal to $p_0 \in (0, 1)$. Then

$$\lim_{t \rightarrow \infty} S^*(t) = \lim_{t \rightarrow \infty} \{1 - [1 - S(t)]^r\}^u = \left\{1 - \left[1 - \lim_{t \rightarrow \infty} S(t)\right]^r\right\}^u = \{1 - [1 - p_0]^r\}^u = p \in (0, 1) \quad (11)$$

Since $0 < 1 - p_0 < 1$, it is easy to see that the last expression in (11) takes a value in $(0, 1)$. The proof is complete. $\square$

We now propose two new defective distributions: the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions.

2.3 The Kumaraswamy Gompertz distribution

Substituting (1) and (2) into (7), (9) and (10), we obtain the Kumaraswamy Gompertz distribution specified by

$$g^*(t) = urbe^{at} \left(e^{\frac{b-hc^{at}}{a}}\right) \left[1 - \left(e^{\frac{b-hc^{at}}{a}}\right)\right]^{r-1} \left[1 - \left\{1 - \left(e^{\frac{b-hc^{at}}{a}}\right)\right\}^r\right]^{u-1}$$

$$S^*(t) = \left\{1 - \left[1 - \left(e^{\frac{b-hc^{at}}{a}}\right)\right]^r\right\}^u$$

and

$$h^*(t) = urbe^{at} \left(e^{\frac{b-hc^{at}}{a}}\right) \left[1 - \left(e^{\frac{b-hc^{at}}{a}}\right)\right]^{r-1} \left[1 - \left\{1 - \left(e^{\frac{b-hc^{at}}{a}}\right)\right\}^r\right]^{u-1}$$

for $a > 0$, $b > 0$, $r > 0$, $u > 0$ and $t > 0$. Figure 3 illustrates possible shapes of these functions.

Figure 3. Probability density, survival and hazard functions of the defective Kumaraswamy Gompertz distribution.

The Kumaraswamy Gompertz distribution is defective if $a < 0$. According to Theorem 2.1, its cure fraction is

$$\lim_{t \rightarrow \infty} S^*(t) = \{1 - [1 - p_0]^r\}^u = p$$

where p_0 is the cure fraction of the defective Gompertz distribution.

2.4 The Kumaraswamy inverse Gaussian distribution

Substituting (4) and (5) into (7), (9) and (10), we obtain the Kumaraswamy inverse Gaussian distribution specified by

$$g^*(t) = \left(\frac{ur}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{(1-at)^2}{2bt} \right\} \right) \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^{r-1} \\ \times \left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^{u-1} \\ S^*(t) = \left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^u$$

and

$$h^*(t) = \left(\frac{ur}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{(1-at)^2}{2bt} \right\} \right) \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^{r-1} \\ \times \left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^{-1}$$

for $a > 0$, $b > 0$, $r > 0$, $u > 0$ and $t > 0$. Figure 4 illustrates possible shapes of these functions.

The Kumaraswamy inverse Gaussian distribution is defective if $a < 0$. According to Theorem 2.1, its cure fraction is

$$\lim_{t \rightarrow \infty} S^*(t) = \{1 - [1 - p_0]^r\}^u = p$$

where p_0 is the cure fraction of the defective inverse Gaussian distribution.

Therefore, now we have two new distributions that can assume a defective form. They have a lot more flexibility and capacity to model different kinds of data sets, at the cost of having two extra parameters.

2.5 Inference

Here, we discuss estimation issues. Consider a data set $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta})$, where $\mathbf{t} = (t_1, \dots, t_n)^T$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$ are the censored failure times. The δ_i is equal to 1 if a failure is observed and 0 otherwise.

Suppose that the data are independently and identically distributed and come from a distribution with probability density and cumulative distribution functions specified by $f(\cdot, \boldsymbol{\theta})$ and $F(\cdot, \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^T$ denotes a vector of parameters. The likelihood function of $\boldsymbol{\theta}$ can be written as²¹

$$L(\boldsymbol{\theta}, \mathbf{D}) \propto \prod_{i=1}^n [f(t_i, \boldsymbol{\theta})^{\delta_i} [1 - F(t_i, \boldsymbol{\theta})]^{1-\delta_i}]$$

Figure 4. Probability density, survival and hazard functions of the defective Kumaraswamy inverse Gaussian distribution.

The corresponding log-likelihood function is

$$\log L(\theta, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \log f(t_i, \theta) + \sum_{i=1}^n (1 - \delta_i) \log [1 - F(t_i, \theta)]$$

For the Kumaraswamy Gompertz distribution,

$$\begin{aligned}\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \left[\log \left(ur b e^{at} \left(e^{\frac{b-b_0 at}{a}} \right) \left[\left(e^{\frac{b-b_0 at}{a}} \right) \right]^{r-1} \left[1 - \left\{ 1 - \left(e^{\frac{b-b_0 at}{a}} \right) \right\}^r \right]^{u-1} \right) \right] \\ + \sum_{i=1}^n (1 - \delta_i) \left[\log \left(\left\{ 1 - \left[1 - \left(e^{\frac{b-b_0 at}{a}} \right) \right]^r \right\}^u \right) \right]\end{aligned}$$

For the Kumaraswamy inverse Gaussian distribution,

$$\begin{aligned}\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \left[\log \left(\left(\frac{ur}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{(1-at)^2}{2bt} \right\} \right) \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^{r-1} \right) \right] \\ + \sum_{i=1}^n \delta_i \left[\log \left(\left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^{u-1} \right) \right] \\ + \sum_{i=1}^n (1 - \delta_i) \left[\log \left(\left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^u \right) \right]\end{aligned}$$

The log-likelihoods can be maximized numerically to obtain the maximum likelihood estimates. There are various routines available for numerical maximization. We used the routine `optim` in the R software²² for numerical maximization. The method of maximization was chosen to be “BFGS”, see the manual for `optim` for details.

Confidence intervals for the parameters were based on asymptotic normality. If $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimator of $\boldsymbol{\theta}$ then it is well known that the distribution of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ can be approximated by a q -variate normal distribution with zero means and covariance matrix $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$, where $\mathbf{I}(\boldsymbol{\theta})$ denotes the observed information matrix defined by

$$\mathbf{I}(\boldsymbol{\theta}) = - \begin{pmatrix} \frac{\partial^2 \log L}{\partial \theta_1^2} & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_q} \\ \frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_q} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L}{\partial \theta_q \partial \theta_1} & \frac{\partial^2 \log L}{\partial \theta_q \partial \theta_2} & \cdots & \frac{\partial^2 \log L}{\partial \theta_q^2} \end{pmatrix}$$

So, an approximate $100(1 - \alpha)$ percent confidence interval for θ_i is $(\hat{\theta}_i - z_{\alpha/2} \sqrt{I^{ii}}, \hat{\theta}_i + z_{\alpha/2} \sqrt{I^{ii}})$, where I^{ii} denotes the i th diagonal element of the inverse of $\mathbf{I}$ evaluated at $\hat{\boldsymbol{\theta}}$ and z_α denotes the $100(1 - \alpha)$ percentile of the standard normal random variable.

2.6 The Kumaraswamy-G regression model

Here, we briefly discuss a possible approach to use the Kumaraswamy family with covariate information. Suppose that $\mathbf{x}' = (1, x_1, \dots, x_p)$ is a vector of covariates from a data set and $S(t)$ is

the survival function of a baseline distribution. Then, the Kumaraswamy- G regression model is given by

$$S^*(t|\mathbf{x}) = \{S^*(t)\}^{\exp(\beta'\mathbf{x})} = \{1 - [1 - S(t)]^r\}^{u\exp(\beta'\mathbf{x})} \quad (12)$$

for $t > 0$, $r > 0$, $u > 0$ and $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ a vector of regression coefficients.

This approach has a simple interpretation when the Kumaraswamy- G distribution is used as a defective model. Suppose that p is the cure fraction of a Kumaraswamy- G defective model, then the limit of the survival function is

$$\lim_{t \rightarrow \infty} S^*(t|\mathbf{x}) = p^{\exp(\beta'\mathbf{x})} \quad (13)$$

It is easy to check that the cured proportion increases when $\beta'\mathbf{x} < 0$ ($\exp(\beta'\mathbf{x}) < 1$) and decreases when $\beta'\mathbf{x} > 0$ ($\exp(\beta'\mathbf{x}) > 1$). Therefore, negative coefficients in β contribute to increase the cure fraction and positive coefficients contribute to decrease it.

The survival function of the Kumaraswamy Gompertz regression model is given by

$$S^*(t|\mathbf{x}) = \left\{1 - \left[1 - \left(e^{\frac{b-at}{a}}\right)^r\right]^r\right\}^{u\exp(\beta'\mathbf{x})} \quad (14)$$

The survival function of the Kumaraswamy inverse Gaussian regression model is given by

$$S^*(t|\mathbf{x}) = \left\{1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b}\Phi\left(\frac{-1-at}{\sqrt{bt}}\right)\right]^r\right\}^{u\exp(\beta'\mathbf{x})} \quad (15)$$

An application of these models to a melanoma data set is described in Section 4.1.

3 Simulation studies

Here, we assess the performance of the maximum likelihood estimates with respect to sample size to show, among other things, that the usual asymptotes of maximum likelihood estimators still hold for defective distributions. The assessment is based on simulations. The description of data generation and details of the distributions simulated from are described below. All computations were performed in R.²²

Suppose that the time of occurrence of an event of interest has cumulative distribution function $F(t)$. We want to simulate a random sample of size n containing real times, censored times and a cure fraction of p . An algorithm for this purpose is:

- determine the desired parameter values, as well as the value of the cure fraction p ;
- generate $M_i \sim \text{Bernoulli}(1-p)$;
- if $M_i = 0$ set $t_i = \infty$; if $M_i = 1$ take t_i as the root of $F(t) = u$, where $u \sim \text{uniform}(0, 1-p)$;
- generate $u'_i \sim \text{uniform}(0, \max(t_i))$, considering only the finite t_i ;
- calculate $t_i = \min(t_i, u'_i)$; if $t_i < u'_i$ set $\delta_i = 1$, otherwise set $\delta_i = 0$.

Note that the range of $F(t)$ has been changed and some adjustments made. Instead of $(0, 1)$, we have used $(0, 1-p)$. Therefore, in the third step of the algorithm, the root of $F(t) - u = 0$ must be for $u \sim \text{uniform}(0, 1-p)$. In the fourth step, the censoring distribution chosen is a uniform

$(0, \max(t_i))$. The limit $\max(t_i)$ was taken in order to control the censoring regardless of the initial parameter choices. In this way, the censoring rates were kept reasonable, as described above.

Tables 1 and 2 describe the results for three different parameters values. The parameter values were selected in order to show the simulation results for small, medium and large cure rates. We took the sample size as $n = 100, 250, 500, 1000, 2000, 5000$. Each sample was replicated 1000 times. In each replication, we computed the maximum likelihood estimates of the parameters, maximum likelihood estimate of the cure fraction, and the standard deviation of the cure fraction (obtained using the delta method). The averages of these estimates over the 1000 replications are reported in Tables 1 and 2. The tables also report the censoring rates in addition to the sample size. Note that censoring rates decrease as sample size increases, but they are still in an appropriate range considering the cure fraction (not too low when the cure fraction is small and not too high when the cure fraction is large).

The following can be observed from the tables:

- the biases and standard deviations generally decrease as sample size increases;
- large sample sizes are needed for the biases of a, b, r, u to become smaller than 10^{-2} , the biases of b and r do not become smaller than 10^{-2} even for $n = 5000$;
- the biases of p become smaller than 10^{-3} for much smaller sample sizes;

Table 1. Simulation of the maximum likelihood estimates for mean and standard deviation of the Kumaraswamy Gompertz distribution.

Sample sizes (Censoring rates)	Mean estimates					Standard deviation estimates				
	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{p}$	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{p}$
$(a, b, r, u, p) = (-1, 5, 5, 0.5, 0.1823)$										
100 (35.80%)	-0.9887	7.0934	8.3434	0.7775	0.1697	0.444	6.1046	7.3722	2.5008	0.2493
250 (33.28%)	-1.0001	6.0915	6.761	0.8496	0.1787	0.2449	3.6785	3.8297	1.6763	0.186
500 (31.36%)	-0.9994	5.6122	5.8102	0.7077	0.1809	0.1571	2.4126	2.0546	0.8084	0.1499
1000 (29.89%)	-1.0076	5.1474	5.2758	0.636	0.1819	0.1045	1.5693	1.1801	0.43	0.1234
2000 (28.92%)	-1.0037	5.101	5.1547	0.5495	0.1824	0.0703	1.0868	0.7892	0.2142	0.1023
5000 (27.65%)	-1.0009	5.0181	5.042	0.5195	0.1822	0.0419	0.6662	0.4727	0.1153	0.0799
$(a, b, r, u, p) = (-2, 2, 2, 2, 0.3605)$										
100 (50.68%)	-1.9845	6.4412	3.1071	1.6974	0.3506	1.0041	13.1274	2.3258	13.5832	0.2808
250 (48.40%)	-1.8958	6.6877	2.8202	1.3912	0.356	0.5	8.3259	1.2862	6.1321	0.2035
500 (47.21%)	-1.8916	5.2443	2.4304	1.5544	0.3579	0.3246	5.4028	0.6895	4.5243	0.1663
1000 (46.16%)	-1.9455	3.7424	2.2102	2.2102	0.3597	0.2323	3.4466	0.4052	5.0514	0.1361
2000 (45.11%)	-1.9719	2.7868	2.0903	2.374	0.3598	0.148	2.0106	0.2358	3.4315	0.1122
5000 (44.16%)	-1.9881	2.3317	2.0379	2.3209	0.3604	0.0943	1.264	0.1468	2.1978	0.0879
$(a, b, r, u, p) = (-3, 11, 2, 0.2, 0.5503)$										
100 (66.01%)	-2.9369	24.1308	3.9845	0.156	0.5254	1.572	58.2769	5.3884	0.9326	0.2893
250 (64.09%)	-3.0368	20.6565	2.99	0.2086	0.5457	0.8143	26.4515	1.7769	0.7126	0.2031
500 (62.89%)	-2.9784	19.0204	2.5784	0.209	0.5479	0.5328	19.116	1.246	0.5011	0.1667
1000 (62.29%)	-2.9769	16.7299	2.3478	0.2251	0.5502	0.3483	12.936	0.6584	0.399	0.1363
2000 (61.46%)	-2.9955	13.4688	2.1343	0.2534	0.5504	0.237	8.4298	0.3923	0.3186	0.1128
5000 (60.71%)	-3.0013	11.6581	2.0375	0.2321	0.5502	0.1463	5.0363	0.2252	0.166	0.0889

Table 2. Simulation of the maximum likelihood estimates for mean and standard deviation of the Kumaraswamy inverse Gaussian distribution.

Sample sizes (Censoring rates)	Mean estimates					Standard deviation estimates				
	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{p}$	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{p}$
$(a,b,r,u,p) = (-1,3,1,1,0.2834)$										
100 (34.38%)	-0.491	4.1373	1.0232	0.9545	0.2771	1.033	13.598	2.6446	1.0137	0.2759
250 (32.53%)	-0.647	6.4553	1.4828	1.055	0.2811	0.8367	13.587	2.3245	0.6701	0.193
500 (31.48%)	-0.632	6.1289	1.4636	1.0436	0.2819	0.5705	9.1473	1.6305	0.4533	0.1514
1000 (31.01%)	-0.555	4.3093	1.1954	1.0104	0.2835	0.3652	4.602	1.0116	0.3298	0.1245
2000 (30.48%)	-0.534	3.7087	1.1205	1.0092	0.2832	0.2406	2.733	0.6694	0.2239	0.1034
5000 (30.10%)	-0.504	3.1797	1.022	0.9942	0.2836	0.139	1.4721	0.3886	0.133	0.0815
$(a,b,r,u,p) = (-0.25,1,2,2,0.3996)$										
100 (47.56%)	-0.233	1.4029	1.9374	2.3161	0.3897	0.5344	5.4392	5.6379	4.9044	0.2851
250 (45.71%)	-0.289	1.9843	2.5624	2.3638	0.3983	0.4737	6.0378	5.2413	2.9732	0.1967
500 (44.60%)	-0.315	2.2856	2.8774	2.2603	0.3985	0.3737	5.2263	4.2558	1.8689	0.1585
1000 (43.67%)	-0.293	1.8463	2.5862	2.1577	0.3981	0.2275	2.6732	2.5578	1.2013	0.1304
2000 (43.19%)	-0.282	1.5384	2.4238	2.1246	0.3991	0.1508	1.5268	1.6917	0.8215	0.1084
5000 (42.66%)	-0.264	1.1796	2.1783	2.0629	0.3994	0.0831	0.6675	0.9237	0.4833	0.0854
$(a,b,r,u,p) = (-1,10,2,0.5,0.5741)$										
100 (62.00%)	-0.629	5.6421	1.0865	0.4459	0.5608	1.7284	24.336	3.6897	0.6395	0.3045
250 (60.97%)	-0.744	8.4841	1.4547	0.4219	0.5694	1.217	22.068	2.8561	0.3205	0.2099
500 (59.86%)	-0.975	12.473	2.0029	0.4701	0.5718	1.075	23.028	2.6607	0.228	0.158
1000 (59.51%)	-1.095	14.577	2.2936	0.4948	0.5735	0.851	19.398	2.1499	0.1635	0.1296
2000 (59.18%)	-1.077	13.452	2.2451	0.4961	0.5744	0.5836	12.844	1.5027	0.1097	0.1077
5000 (58.87%)	-1.044	11.539	2.1227	0.5002	0.5747	0.3412	6.7388	0.8807	0.0657	0.085

- although the standard deviations decrease as n increases, their values are not so small even for n as large as 5000. Sample sizes larger than 5000 are not realistic in practical survival analysis.

These observations are for specific parameter values and $n = 100, 250, 1000, 2000, 5000$. But the same observations held for a wide range of other parameter values and a wide range of the values of $n < 5000$.

4 Applications

We illustrate the distributions in Section 2 using three real cancer data sets. The data sets have different sample sizes and exhibit distinct Kaplan–Meier curves.²³

The following distributions were fitted to each of the data sets: the Gompertz distribution, the inverse Gaussian distribution, the Kumaraswamy Gompertz distribution and the Kumaraswamy inverse Gaussian distribution. This allows use to see whether the Kumaraswamy Gompertz distribution provides a better fit than the Gompertz distribution or whether the Kumaraswamy inverse Gaussian distribution provides a better fit than the inverse Gaussian distribution.

For each fitted distribution, we provide the maximum likelihood estimates, their standard errors (in parentheses), the values of the AIC (Akaike information criterion), the values of the BIC (Bayesian information criterion) and the values of the CAIC (consistent Akaike information criterion). All calculations were performed using the R software.²² As discussed in the

methodology section, the point estimate of p was obtained using the estimates of the other parameters. The standard deviation of the estimator of p was calculated using the delta method with a first-order Taylor's approximation.

An estimate of the survival times was also obtained using the Kaplan–Meier estimator. As an experiment comes to an end, some of the observed elements do not fail. However, it is not known whether these elements would fail at some given point, after the end of the observed period. Depending on what is being analyzed, there are two possibilities; that the elements will never fail, or will fail at some point in the future if given enough time. The end of a Kaplan–Meier curve is an estimate of the cure fraction, with the assumption that elements in the study which did not fail, will never fail at all at any point in the future, after the end of the study. It should be noted that this is a difficult assumption to make.

We also compared the parametric curves with the Kaplan–Meier curve. Closer they are to each other, the better the fit. The comparison was a visual as well as a numerical assessment. For the latter, the sum of absolute deviations was used.

4.1 Melanoma cancer data

This data set collected in the period 1991–1998 is related to a clinical study in which patients were observed for recurrence after a removal of a malignant melanoma. Melanoma is a type of cancer that develops in melanocytes, responsible for skin pigmentation. This is a potentially serious malignant tumor that may arise in the skin, mucous membranes, eyes and central nervous system, with a great risk of producing metastases and high mortality rates in the later stages. The data set has 417 observations with 55.63 per cent of censored elements (232 in total). The maximum observation time is 7 years and the patient's death is considered the event of interest. A detailed description of the data can be found in Ibrahim et al.²⁴

The survival curves of the fitted distributions are shown in Figure 5. The corresponding maximum likelihood estimates are given in Table 3.

The estimated value of a is negative for all of the distributions, so all of the fitted distributions are defective. The cure fraction estimates for the Gompertz and inverse Gaussian distributions are 25.56 and 13.98 per cent, respectively. These estimates are smaller than the end of the Kaplan–Meier curve, which stabilizes at 48.04 per cent. The cure fraction estimates for the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are 49.01 and 41.46 per cent, respectively.

The survival curves of the fitted Kumaraswamy distributions capture the Kaplan–Meier curve more accurately than the survival curves of the fitted baseline distributions. This is especially the case for the fitted Kumaraswamy Gompertz distribution.

We can also see that the AIC and CAIC values are smaller for the Kumaraswamy distributions than for the baseline distributions. The AIC and CAIC values for the inverse Gaussian distribution are a little larger than those for the Kumaraswamy inverse Gaussian distribution. Both of these distributions outperform the Gompertz distribution, but do not perform as well as the Kumaraswamy Gompertz distribution. In terms of BIC, the best fit is given by the inverse Gaussian distribution, followed by the Kumaraswamy Gompertz distribution, then the Kumaraswamy inverse Gaussian distribution and then the Gompertz distribution.

Now we present an application of the regression model described in Section 2.6. We are going to consider a covariate that represents the age of the individuals in the data set. For a simple illustrative example, we categorize this variable into two classes. The variable was classified as zero when the age was below the average of all individuals and as one otherwise.

Figure 5. Survival curves for the fitted distributions for the melanoma data set.

Table 3. Maximum likelihood estimates for the fitted distributions for the melanoma data set.

Model	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{p}$	AIC	BIC	CAIC
Gompertz	-0.1314 (0.0540)	0.1793 (0.0217)	1 —	1 —	0.2556 (0.3359)	1096.467	1104.534	1096.496
Kumaraswamy Gompertz	-0.5929 (0.1441)	0.6040 (0.9382)	2.8708 (0.7902)	2.2039 (6.6481)	0.4901 (0.1872)	1057.318	1073.45	1057.415
Inverse Gaussian	-0.0357 (0.0319)	0.4740 (0.0427)	1 —	1 —	0.1398 (0.1084)	1062.956	1071.022	1062.985
Kumaraswamy Inverse Gaussian	-2.5544 (0.8850)	32.6713 (18.6450)	22.0545 (7.5425)	27.2590 (18.2056)	0.4146 (0.0475)	1061.043	1077.175	1061.14

Table 4 reports the maximum likelihood estimates of the proposed regression models. We can see that the Gompertz and inverse Gaussian distributions estimate the cure fraction of both groups very closely. The cure fraction for the group with age below the average is around 0.28. That for the group with age above the average is 0.22. The Kumaraswamy inverse Gaussian distribution gives higher estimates, 0.34 and 0.28. The Kumaraswamy Gompertz distribution also gives higher estimates, 0.52 and 0.46.

Figures 6 and 7 show the fitted survival curves of the baseline and Kumaraswamy distributions together with Kaplan–Meier curves, plotted in the same color. The dashed lines represent the estimated cure fractions. As we can see, the baseline distributions do not capture the Kaplan–Meier curve as well as the Kumaraswamy distributions do. Kumaraswamy distributions fit the end of the curve better, leading to higher cure fraction estimates. The baseline distributions give more conservative estimates.

Table 4. Maximum likelihood estimates for the fitted regression models for the melanoma data set.

Model	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}_0$	$\hat{\rho}_1$
Gompertz	−0.8988 (0.3799)	0.0264 (0.0388)	− −	− −	3.7657 (1.4764)	0.1893 (0.1472)	0.2812 (0.3480)	0.2158 (0.3380)
Kumaraswamy	−4.3401 (0.8070)	2.6954 (3.4839)	2.7143 (0.5119)	0.6281 (1.3449)	2.0631 (3.4039)	0.1719 (0.1472)	0.5215 (0.2017)	0.4616 (0.2075)
Inverse Gaussian	−0.6542 (1.0402)	3.1659 (0.4712)	− −	− −	0.1386 (0.6018)	0.1789 (0.1472)	0.2881 (0.4972)	0.2258 (0.4832)
Kumaraswamy	−0.8113 (1.8714)	2.3698 (4.2916)	0.8062 (1.3570)	3.3050 (7.1690)	−0.9812 (2.5830)	0.1753 (0.1473)	0.3456 (0.4744)	0.2819 (0.4697)

Figure 6. The fitted Gompertz regression model (left) and the inverse Gaussian model (right).

In all fits, the group consisting of younger individuals has a survival probability higher than the group consisting of older individuals. That makes sense and it is fair to expect this behavior in all data sets of the kind.

4.2 Colon cancer data

The second data set arises from one of the first successful trials of adjuvant chemotherapy for colon cancer. The event of interest here is the recurrence or death for the individual under the proposed treatment. The data set has 1858 observations with 50.58 per cent censoring (938 in total). The data is available in R²² in the survival package. Details of the data can be found in Laurie et al.²⁵ For computational purposes, the observed times were divided by their maximum, putting all observations in the (0, 1) range. Multiplying the data values by a constant does not change anything of interest for this paper.

Figure 7. The fitted Kumaraswamy Gompertz regression model (left) and the Kumaraswamy inverse Gaussian model (right).

Figure 8. Survival curves for the fitted distributions for the colon data set.

The survival curves of the fitted distributions are shown in Figure 8. The corresponding maximum likelihood estimates are given in Table 5.

Again, the estimated value of a is negative for all of the distributions, and so all of the fitted distributions are defective. The cure fraction estimates for the Gompertz and inverse Gaussian distributions are 42.47 and 36.53 per cent, respectively. The cure fraction estimates for the

Table 5. Maximum likelihood estimates for the fitted distributions for the colon data set.

Model	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{p}$	AIC	BIC	CAIC
Gompertz	-2.3375 (0.1772)	2.0018 (0.1025)	1 —	1 —	0.4247 (0.1339)	1518.022	1529.076	1518.028
Kumaraswamy Gompertz	-3.3598 (0.2727)	17.3085 (6.4465)	2.0508 (0.2476)	0.1757 (0.0803)	0.4586 (0.1215)	1452.173	1474.282	1452.195
Inverse Gaussian	-1.6688 (0.1568)	7.3406 (0.2901)	1 —	1 —	0.3653 (0.0219)	1597.472	1608.527	1597.479
Kumaraswamy inverse Gaussian	-1.3366 (2.861)	91.5128 (29.5785)	5.6371 (1.3563)	0.9717 (0.3392)	0.1602 (0.2075)	1511.714	1533.823	1511.735

Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are 45.86 and 16.02 percent, respectively. The cure fraction estimates for the inverse Gaussian and Kumaraswamy inverse Gaussian distributions are smaller than the end of the Kaplan–Meier curve, which stabilizes at 46.51 per cent. The cure fraction estimates for the Gompertz and Kumaraswamy Gompertz distributions are a lot closer to the end of the Kaplan–Meier curve.

The survival curves of the fitted Kumaraswamy distributions capture the Kaplan–Meier curve more accurately than the survival curves of the fitted baseline distributions. However, it must be said that all four distributions provide reasonable fits.

The AIC and CAIC values show that the Kumaraswamy Gompertz distribution gives the best fit by far. The second smallest values for AIC and CAIC are for the Kumaraswamy inverse Gaussian distribution. The third smallest values (a little larger than the second smallest values) for AIC and CAIC are for the Gompertz distribution. The largest values (much larger than the third smallest values) for AIC and CAIC are for the inverse Gaussian distribution. The BIC values show that the Kumaraswamy Gompertz distribution gives the best fit, followed by the Gompertz distribution, then the Kumaraswamy inverse Gaussian distribution and then the inverse Gaussian distribution.

4.3 Leukemia data

The third data set relates to a study of the recurrence of leukemia in patients who were submitted to a certain kind of transplantation. Leukemia is a type of cancer that affects the white blood cells produced by the bone marrow and can take several forms. The data set has 44 observations with 20.45 per cent censoring (9 in total). The maximum observation time was approximately 5 years. For details of the data, see Kersey et al.²⁶

The survival curves of the fitted distributions are shown in Figure 9. The corresponding maximum likelihood estimates are given in Table 6.

Not all of the fitted distributions estimate a by a negative value. The inverse Gaussian distribution is not estimated as being defective, leading to an estimated cure rate of 0. The cure fraction estimate for the Gompertz distribution is 20.73 per cent. The cure fraction estimates for the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are 19.61 and 20.25 per cent, respectively. The cure fraction estimates for the Gompertz, Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are close to the end of Kaplan–Meier curve, which stabilizes at 19.88 per cent.

The survival curves of the fitted Kumaraswamy distributions capture the Kaplan–Meier curve more accurately than the survival curves of the fitted baseline distributions. However, it must be said

Figure 9. Survival curves for the fitted distributions for the leukemia data set.

Table 6. Maximum likelihood estimates for the fitted distributions for the leukemia data set.

Model	$\hat{a}$	$\hat{b}$	$\hat{r}$	$\hat{u}$	$\hat{p}$	AIC	BIC	CAIC
Gompertz	-1.5103 (0.3696)	2.3767 (0.5171)	1 —	1 —	0.2073 (0.2557)	52.5763	56.1447	52.869
Kumaraswamy Gompertz	-2.9825 (0.9399)	12.1742 (11.5263)	7.1974 (6.0582)	0.7541 (1.3356)	0.1961 (0.2448)	36.4139	43.5507	37.4396
Inverse Gaussian	0.2261 (0.3436)	3.1393 (0.7379)	1 —	1 —	0 —	50.9801	54.5485	51.2728
Kumaraswamy inverse Gaussian	-1.6449 (1.2734)	0.8895 (1.6488)	0.7226 (1.2366)	22.3123 (42.1862)	0.2025 (0.0612)	36.5291	43.6659	37.5548

that the Gompertz distribution provides a reasonable fit as well. The inverse Gaussian distribution does not fit well.

The AIC, BIC and CAIC values show that the Kumaraswamy Gompertz distribution gives the best fit. Gompertz distribution gives the largest values for AIC, BIC, CAIC in spite of its good fit to the Kaplan–Meier curve. This appears a little strange. But closeness of the fitted survival curves to the Kaplan–Meier curve is only a “measure” just like the AIC, BIC and CAIC. It just happens that the inverse Gaussian distribution does not perform so well with respect to this measure but the Gompertz distribution does so.

4.4 Discussion

Figure 10 plots the observed probabilities versus the expected probabilities for the fitted Gompertz, inverse Gaussian, Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions for the three data sets. We see that the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian

Figure 10. Probability plots for the fit of the four distributions to the three data sets.

distributions have the points closer to the diagonal line for each of the data sets (especially when the Kumaraswamy Gompertz and Gompertz distributions are compared for the melanoma and leukemia data sets and when the Kumaraswamy inverse Gaussian and inverse Gaussian distributions are compared for the colon and leukemia data sets), showing their better fits.

Table 7 reports the statistic and p -value of the log-likelihood ratio test. The test compares the baseline distribution with the corresponding Kumaraswamy distribution. The p -values are very small except for the Kumaraswamy inverse Gaussian distribution with a p -value of 0.052. These are compatible with the visual assessment of the survival curves in Figure 10.

Table 8 reports the 95 per cent asymptotic confidence intervals for the parameter a . The objective here is to see whether the cure rate p is different from zero. If the parameter a is significantly negative, we can say that p is significantly different from zero. This is because a is free to take any value on the real line, while the other parameters are positive by definition. The Gompertz and Kumaraswamy Gompertz distributions have fully negative intervals for all data sets. The inverse Gaussian distribution has a fully negative interval for only the colon data set. The Kumaraswamy inverse Gaussian distribution has a fully negative interval for only the melanoma data set. Therefore, the cure fraction is significantly different from zero in most cases. For every data set, there are at least two distributions indicating that the cure fraction p is significantly different from zero.

The AIC values for the fitted distributions and those for the standard mixture model are compared in Table 9. The defective model was compared with its own version as a standard

Table 7. Log-likelihood ratio test for the proposed models and data sets.

Data set	Model	LR test statistic	p-value
Melanoma	Gompertz/Kum–Gompertz	43.149	<0.0001
	Inverse Gaussian/Kum–inverse Gaussian	5.913	0.052
Colon	Gompertz/Kum–Gompertz	69.849	<0.0001
	Inverse Gaussian/Kum–inverse Gaussian	89.758	<0.0001
Leukemia	Gompertz/Kum–Gompertz	20.1624	<0.0001
	Inverse Gaussian/Kum–inverse Gaussian	18.451	0.0001

Table 8. The 95 per cent asymptotic confidence intervals for the parameter a for the proposed models and data sets.

Model	Melanoma	Colon	Leukemia
Gompertz	(−0.2372; −0.0256)	(−2.6848; −1.9902)	(−2.2347; −0.7859)
Kumaraswamy Gompertz	(−0.8753; −0.3105)	(−3.8943; −2.8253)	(−4.8247; −1.1403)
Inverse Gaussian	(−0.0982; 0.0268)	(−1.9761; −1.3615)	(−0.4474; 0.8996)
Kumaraswamy inverse Gaussian	(−4.2890; −0.8198)	(−6.9442; 4.2710)	(−4.1408; 0.8510)

Table 9. AIC values for the fitted distributions and for the standard mixture model. The smaller AIC values are shown in bold.

Model	Melanoma		Colon		Leukemia	
	Defective	Mixture	Defective	Mixture	Defective	Mixture
Gompertz	1096.467	1085.459	1518.022	1520.023	52.5763	50.741
Kumaraswamy Gompertz	1057.318	1058.741	1452.173	1451.823	36.4139	36.432
Inverse Gaussian	1062.956	1063.200	1597.472	1668.362	50.9801	37.139
Kumaraswamy inverse Gaussian	1061.043	1058.741	1511.714	1496.993	36.5291	37.473

mixture model. In this way, the mixture model has always one extra parameter than the defective one. As noted before, we can see that the Kumaraswamy distributions have generally smaller AIC values, in the defective form or in the mixture model form. For the melanoma data set, all but the Gompertz and Kumaraswamy inverse Gaussian distributions have smaller AIC values. For the colon data set, the baseline distributions perform better under the defective strategy while the mixture models have better values in the Kumaraswamy forms. For the leukemia data set, the baseline distributions perform better under the mixture models strategy while the Kumaraswamy distributions are better under the defective strategy.

We have presented three different examples on defective models. The examples show that there are cases where the baseline distributions perform well and some cases where they do not. The examples also show that the Kumaraswamy distributions lead to better fits in most cases. We also see that there are cases where the defective models outperform the mixture models.

5 Conclusions

We have proposed two new defective distributions: the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions. Each of these distributions has one less parameter than its version under the standard mixture approach. We have assessed the finite sample performance of their maximum likelihood estimators. We have illustrated the use of the new distributions to three real cancer data sets containing cure fractions. We have shown that the new distributions perform better than the baseline distributions (the Gompertz and inverse Gaussian distributions) in terms of the AIC, BIC, CAIC and proximity to the Kaplan–Meier curves. In some cases, they outperform the standard mixture model in terms of AIC. We have also illustrated the use of regression models based on the new distributions to one of the three data sets.

A future work is to compare the proposed distributions to semiparametric distributions for cure rate modeling such as piecewise exponential distributions.

Acknowledgments

The authors thank CNPq and FAPESP for financial support during the course of this project. The authors also thank the editor and the two referees for careful reading and comments which greatly improved the paper.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

- Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Statist Soc Ser B* 1949; **11**(1): 15–53.
- Berkson J and Gage RP. Survival curve for cancer patients following treatment. *J Am Statist Assoc* 1952; **47**(259): 501–515.
- Yakovlev AY and Tsiodikov AD. *Stochastic models of tumor latency and their biostatistical applications*. Vol. 1, Singapore: World Scientific, 1996.
- Chen MH, Ibrahim JG and Sinha D. A new Bayesian model for survival data with a surviving fraction. *J Am Statist Assoc* 1999; **94**(447): 909–919.
- Ibrahim JG, Chen MH and Sinha D. *Bayesian Survival Analysis*. Wiley Online Library, 2005.
- Balka J, Desmond AF and McNicholas PD. Review and implementation of cure models based on first hitting times for wiener processes. *Lifetime Data Anal* 2009; **15**(2): 147–176.
- Balka J, Desmond AF and McNicholas PD. Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *J Appl Statist* 2011; **38**(1): 127–144.
- Gieser PW, Chang MN, Rao PV, et al. Modelling cure rates using the Gompertz model with covariate information. *Statist Med* 1998; **17**(8): 831–839.
- Cordeiro GM and de Castro M. A new family of generalized distributions. *J Statist Computat Sim* 2011; **81**(7): 883–898.
- Haybittle JL. The estimation of the proportion of patients cured after treatment for cancer of the breast. *Br J Radiol* 1959; **32**(383): 725–733.
- Cantor AB and Shuster JJ. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statist Med* 1992; **11**(7): 931–937.
- Schrödinger E. Zur theorie der fall-und steigversuche an teilchen mit brownscher bewegung. *Phys Zeitschrift* 1915; **16**: 289–295.
- Tweedie MCK. Inverse statistical variates. *Nature* 1945; **155**(3937): 453.
- Kumaraswamy P. A generalized probability density function for double-bounded random processes. *J Hydrol* 1980; **46**(1): 79–88.
- Cordeiro GM, Nadarajah S and Ortega EMM. The Kumaraswamy Gumbel distribution. *Statist Meth Applic* 2012; **21**(2): 139–168.
- Cordeiro GM, Ortega EMM and Nadarajah S. The Kumaraswamy Weibull distribution with application to failure data. *J Franklin Inst* 2010; **347**(8): 1399–1429.
- Saulo H, Leão J and Bourguignon M. The Kumaraswamy Birnbaum-Saunders distribution. *J Statist Theory Pract* 2012; **6**(4): 745–759.
- de Santana TVF, Ortega EMM, Cordeiro GM, et al. The Kumaraswamy log-logistic distribution. *J Statist Theory Applicat* 2012; **11**(3): 265–291.
- de Pascoa MAR, Ortega EMM and Cordeiro GM. The Kumaraswamy generalized gamma distribution with application in survival analysis. *Statist Methodol* 2011; **8**(5): 411–433.
- Paranaíba PF, Ortega EMM, Cordeiro GM, et al. The Kumaraswamy Burr XII distribution: Theory and practice. *J Statist Computat Sim* 2013; **83**(11): 2117–2143.
- Klein JP and Moeschberger ML. *Survival analysis: Statistical methods for censored and truncated data*. New York: Springer-Verlag, 2003.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. <http://www.R-project.org/>

23. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J Am Statist Assoc* 1958; **53**(282): 457–481.
24. Ibrahim JG, Chen MH and Sinha D. Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* 2001; **57**(2): 383–388.
25. Laurie JA, Moertel CG, Fleming TR, et al. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. *J Clin Oncol* 1989; **7**(10): 1447–1456.
26. Kersey JH, Weisdorf D, Nesbit ME, et al. Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *N Engl J Med* 1987; **317**(8): 461–467.