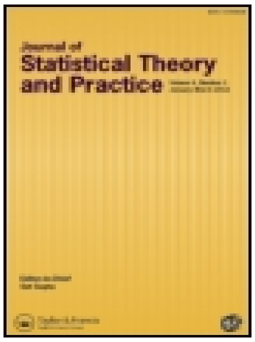

Journal of Statistical Theory and Practice

Journal of Statistical Theory and Practice

ISSN: 1559-8608 (Print) 1559-8616 (Online) Journal homepage: http://www.tandfonline.com/loi/ujsp20

A new straightforward defective distribution for survival analysis in the presence of a cure fraction

Edson Z. Martinez & Jorge A. Achcar

To cite this article: Edson Z. Martinez & Jorge A. Achcar (2018): A new straightforward defective distribution for survival analysis in the presence of a cure fraction, Journal of Statistical Theory and Practice, DOI: 10.1080/15598608.2018.1460885

To link to this article: https://doi.org/10.1080/15598608.2018.1460885

A new straightforward defective distribution for survival analysis in the presence of a cure fraction

Edson Z. Martinez[a], Jorge A. Achcar[a]

[a]*Department of Social Medicine, University of São Paulo (USP), Ribeirão Preto Medical School, Brazil.*

Abstract

In the present article we introduce a new distribution, namely, the defective Dagum distribution (DDD). This improper distribution can be seen as an extension of the Type I Dagum distribution and it is useful to accommodate survival data in the presence of a cure fraction. In the applications of survival methods to medical data, the cure fraction is defined as the proportion of patients who are cured of disease and become long-term survivors. The great advantage of the DDD is that the cure fraction can be written as a function of only one parameter. We also considered the presence of censored data and covariates. Maximum likelihood and Bayesian methods for estimation of the model parameters are presented. A simulation study is provided to evaluate the performance of the maximum likelihood method in estimating parameters. In the Bayesian analysis, posterior distributions of the parameters are estimated using the Markov chain Monte Carlo (MCMC) method. An example involving a real data set is presented. The model based on the new distribution is easy to use and it is a good alternative for the analysis of real time-to-event data in the presence of censored information and a cure fraction.

Keywords: Dagum distribution, Survival analysis, Defective distributions, Bayesian methods, Censored data.

1. Introduction

The usual methods in survival analysis assume that all sampled individuals are susceptible to the event of interest. In mathematical terms this means that the cumulative probability $F(t)$ of occurrence of the event at the time t tends to 1 when t is a number sufficiently large. However, in clinical trials involving particular conditions this assumption can break down if there is a proportion of patients who respond favourably to treatment and they are consequently regarded as cured of their disease. In this case, considering that the event of interest is death due to this specific disease, a Kaplan-Meier plot of the time to this event describes a survival function with a stable plateau at the right end of the curve (13). Common statistical approaches applied to data sets with a proportion of immunes include the mixture model (9), that explicitly includes a parameter accounting for the cure rate. This model assumes that the probability of observing a survival time greater than or equal to some fixed value t is given by the survival function

$$S(t) = 1 - F(t) = P(T > t) = p + (1-p)S_0(t), \tag{1}$$

where p is a parameter which represents the proportion of immune individuals ($0 < p < 1$) and

$S_0(t)$ is the baseline survival function for the susceptible individuals. Usual choices for $S_0(t)$ include the Weibull distribution and its extensions and generalizations. If $S_0(t)$ is a proper survival function, it is straightforward to see that $S(t)$ tends to p as t tends to infinity. There are a significant number of articles in the literature about different extensions and applications of the mixture cure fraction model, using Bayesian and frequentist methods of inference. For example, Chen et al. (5) extended this model to a multivariate case where the correlation structure between the failure times is described by a frailty term, which is assumed to have a positive stable distribution. A bivariate cure-mixture model was also introduced by Wienke et al. (20), considering left-truncated and right-censored lifetime data. Xiang et al. (19) proposed a mixture cure modeling procedure for analyzing clustered and interval-censored survival time data by incorporating random effects in both the regression components.

Models based on defective distributions are alternatives to the mixture models. Defective distributions are characterized as distributions that are not normalized to one for some values of their parameters. These distributions allow us to fit survival data including both immune and susceptible individuals without explicitly including the parameter p. In the literature, usual defective distributions include the defective Gompertz distribution (12), the defective inverse Gaussian distribution (1) and the exponentiated-Weibull distribution (4). A disadvantage of the defective Gompertz distribution is that its hazard function is always decreasing. In another hand, the defective inverse Gaussian distribution has a unimodal hazard function. More recently new defective distributions based on the Kumaraswamy and the Marshall-Olkin families were introduced in the literature (16; 17). These distributions can accommodate more flexible hazard shapes.

In the present study we introduce a new straightforward defective distribution for survival analysis in the presence of a cure fraction, based on the Dagum distribution (6). Let T be a random variable representing the survival time to some event of interest, and let t be an observation of T. The proper three-parameters type I Dagum distribution has cumulative distribution function given by

$$F(t;\boldsymbol{\lambda}) = \left[k + \left(\frac{t}{\beta}\right)^{-\alpha}\right]^{-\gamma}, \qquad (2)$$

where $k=1$, $\alpha>0$ is a shape parameter, $\beta>0$ is a scale parameter, $\gamma>1$ is also a shape parameter and $t>0$ (7; 8). Note that the case $k=1$ and $\gamma=1$ is the log-logistic distribution (18). The distribution proposed in this article considers $k=\theta^{-1}$ and $\gamma=1$, where θ is a parameter related to the proportion of susceptible individuals. The main advantage of the proposed method is the simplicity of the expression of the cure fraction, which only contains one parameter.

2. Methods

2.1. The defective Dagun distribution (DDD)

The proposed defective three-parameters distribution has cumulative distribution function given by

$$F(t;\boldsymbol{\lambda}) = \frac{\theta\beta}{\beta + \theta t^{-\alpha}}, \tag{3}$$

where $t > 0$, $\alpha > 0$, $\beta > 0$, $0 < \theta < 1$ and $\boldsymbol{\lambda} = (\alpha, \beta, \theta)$ is the vector of unknown parameters to be estimated. We will call this new distribution as the defective Dagun distribution (DDD). A random variable T that follows a DDD will be denoted by $T \sim DDD(\alpha, \beta, \theta)$. Considering that $\lim_{t\to\infty} F(t;\boldsymbol{\lambda}) = \theta$, we have that $F(t;\boldsymbol{\lambda})$ only describes a proper distribution if $\theta = 1$. Thus, the parameter $0 < \theta < 1$ is used to account for the presence of a proportion of immune individuals. The cure rate p is then defined as

$$p = \lim_{t\to\infty} S(t;\boldsymbol{\lambda}) = 1 - \theta, \tag{4}$$

$0 < p < 1$, where $S(t;\boldsymbol{\lambda}) = 1 - F(t;\boldsymbol{\lambda})$ is the survival function. In addition, the corresponding probability density function is given by

$$f(t;\boldsymbol{\lambda}) = \alpha\beta\theta^2 \frac{t^{-(\alpha+1)}}{\left(\beta + \theta t^{-\alpha}\right)^2} \tag{5}$$

and the hazard function is

$$h(t;\boldsymbol{\lambda}) = \frac{f(t;\boldsymbol{\lambda})}{S(t;\boldsymbol{\lambda})} = \frac{\alpha\beta\theta^2 t^{-(\alpha+1)}}{\left(\beta + \theta t^{-\alpha}\right)\left(\beta + \theta t^{-\alpha} - \theta\beta\right)}. \tag{6}$$

From the cumulative function (3), we have that the quantiles can be obtained using the expression

$$F^{-1}(u;\boldsymbol{\lambda}) = \left[\frac{u\theta}{\beta(\theta - u)}\right]^{\frac{1}{\alpha}}, \tag{7}$$

where $0 < u < 1$. As, a special case, the median of T is

$$M_e(T) = \left[\frac{\theta}{\beta(2\theta - 1)}\right]^{\frac{1}{\alpha}}. \tag{8}$$

Note that the median does not exist if $\theta \le 0.5$.

Figure 1 gives examples of the survival and hazard functions for the DDD, based on some choices for the parameters α, β and θ. We can note that the hazard function is a decreasing function for $\alpha \le 1$ or a unimodal function for $\alpha > 1$.

Figure 1: Plots of the survival (left panels) and hazard functions (right panels) for the DDD, based on some choices for the parameters α, β and θ.

The likelihood function $L(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$ under right censoring is given by

$$L(\boldsymbol{\lambda})=\prod_{i=1}^{n}\left[f(t_i;\boldsymbol{\lambda})\right]^{d_i}\left[S(t_i;\boldsymbol{\lambda})\right]^{1-d_i}, \tag{9}$$

where d_i is a censoring indicator variable, such that $d_i=1$ if the i th subject is observed until the occurrence of the event and $d_i=0$ otherwise. From the expressions (3) and (5), the likelihood function for the model based on the DDD is given by

$$L(\boldsymbol{\lambda})=\prod_{i=1}^{n}\left[\alpha\beta\theta^2\frac{t_i^{-(\alpha+1)}}{\left(\beta+\theta t_i^{-\alpha}\right)^2}\right]^{d_i}\left[1-\frac{\theta\beta}{\beta+\theta t_i^{-\alpha}}\right]^{1-d_i} \tag{10}$$

and the respective log-likelihood function is

$$l(\boldsymbol{\lambda})=\ln\left(\alpha\beta\theta^2\right)\sum_{i=1}^{n}d_i-(\alpha+1)\sum_{i=1}^{n}d_i\ln t_i-2\sum_{i=1}^{n}d_i\ln\left(\beta+\theta t_i^{-\alpha}\right)+\sum_{i=1}^{n}(1-d_i)\ln\left(1-\frac{\theta\beta}{\beta+\theta t_i^{-\alpha}}\right). \tag{11}$$

By deriving the log-likelihood function with respect to α, β and θ, we have the following equations:

$$\frac{\partial}{\partial\alpha}l(\boldsymbol{\lambda})=\frac{1}{\alpha}\sum_{i=1}^{n}d_i-\sum_{i=1}^{n}d_i\ln t_i+2\theta\sum_{i=1}^{n}\frac{d_i\ln t_i}{\theta+\beta t_i^{\alpha}}-\beta\theta^2\sum_{i=1}^{n}\frac{(1-d_i)\ln t_i}{\left(\beta+\theta t_i^{-\alpha}\right)\left[\theta+\beta t_i^{\alpha}(1-\theta)\right]}, \tag{12}$$

$$\frac{\partial}{\partial\beta}l(\boldsymbol{\lambda})=\frac{1}{\beta}\sum_{i=1}^{n}d_i-2\sum_{i=1}^{n}\frac{d_i}{\beta+\theta t_i^{-\alpha}}-\theta^2\sum_{i=1}^{n}\frac{1-d_i}{\left(\beta+\theta t_i^{-\alpha}\right)\left[\theta+\beta t_i^{\alpha}(1-\theta)\right]} \tag{13}$$

and

$$\frac{\partial}{\partial\theta}l(\boldsymbol{\lambda})=\frac{2}{\theta}\sum_{i=1}^{n}d_i-2\sum_{i=1}^{n}\frac{d_i}{\theta+\beta t_i^{\alpha}}-\beta^2\sum_{i=1}^{n}\frac{1-d_i}{\left(\beta+\theta t_i^{-\alpha}\right)\left[\beta(1-\theta)+\theta t_i^{-\alpha}\right]}. \tag{14}$$

The maximum likelihood estimators (MLEs) are obtained from the numerical maximization of equation (11), since the solution of the maximum likelihood equations is not in closed form. As an alternative, software R provides the package maxLik to solve these equations. Confidence intervals (CI) for α, β and θ can be obtained using the asymptotical normal distribution for the respective estimators α_{ML}, β_{ML} and θ_{ML}, that is,

$$\boldsymbol{\lambda}\overset{a}{\sim}N\left(\boldsymbol{\lambda},\mathbf{I}_0^{-1}\right), \tag{15}$$

where $\mathbf{I}_0$ denotes the observed Fisher information matrix given by

$$\mathbf{I}_0(\boldsymbol{\lambda})=\begin{bmatrix}I_{\alpha\alpha} & I_{\alpha\beta} & I_{\alpha\theta}\\ I_{\alpha\beta} & I_{\beta\beta} & I_{\beta\theta}\\ I_{\alpha\theta} & I_{\beta\theta} & I_{\theta\theta}\end{bmatrix} \tag{16}$$

and I_{jk} denotes the second derivative of the log-likelihood with minus sign which respect to the jth and kth parameters locally at the MLE of the parameters. The inverse of the observed Fisher information matrix I_0 gives the approximate variances and covariances of the MLEs of the parameters. In this way, we can evaluate approximate standard errors of the estimates.

Since $\alpha > 0$, $\beta > 0$ and $0 < \theta < 1$, a convenient reparametrization of the model considers $\ln \alpha = \alpha_0$, $\ln \beta = \beta_0$ and

$$\theta = \frac{e^{\theta_0}}{1+e^{\theta_0}}, \tag{17}$$

where α_0, β_0 and θ_0 are real numbers. In this case, the cure rate p is defined as

$$p = 1-\theta = \frac{1}{1+e^{\theta_0}}, \tag{18}$$

and by the invariance property of the maximum likelihood estimators, Wald-type asymptotic $100(1-\varphi)\%$ confidence intervals for α, β and θ are given by

$$\exp\left[\alpha_{0ML} \mp z_{\varphi/2} \times \sqrt{Var\left(\alpha_{0ML}\right)}\right], \tag{19}$$

$$\exp\left[\beta_{0ML} \mp z_{\varphi/2} \times \sqrt{Var\left(\beta_{0ML}\right)}\right] \tag{20}$$

and

$$1-\left\{1-\exp\left[\theta_{0ML} \mp z_{\varphi/2} \times \sqrt{Var\left(\theta_{0ML}\right)}\right]\right\}^{-1}, \tag{21}$$

respectively, where α_{0ML}, β_{0ML} and θ_{0ML} are the respective maximum likelihood estimators for α_0, β_0 and θ_0, $Var\left(\alpha_{0ML}\right)$, $Var\left(\beta_{0ML}\right)$ and $Var\left(\theta_{0ML}\right)$ are estimates for the variances of α_{0ML}, β_{0ML} and θ_{0ML}, respectively, and $z_{\varphi/2}$ is the upper $(\varphi/2)th$ percentile of a standard normal distribution. This reparametrization is useful especially to avoid obtaining confidence limits for α and β outside the interval $(0,\infty)$ and confidence limits for θ outside the interval $(0,1)$.

2.2. *Model with covariates*

In order to include covariates into the proposed model, the parameters α, β and θ in the likelihood function (10) can be replaced respectively by the functions $\alpha(\mathbf{x}_i)$, $\beta(\mathbf{w}_i)$ and $\theta(\mathbf{z}_i)$, given by

$$\alpha(\mathbf{x}_i) = \exp\left(\mathbf{x}_i\boldsymbol{\alpha}^*\right), \tag{22}$$

$$\beta(\mathbf{w}_i) = \exp\left(\mathbf{w}_i\boldsymbol{\beta}^*\right) \tag{23}$$

and

$$\theta(\mathbf{z}_i) = \frac{\exp\left(\mathbf{z}_i\boldsymbol{\theta}^*\right)}{1+\exp\left(\mathbf{z}_i\boldsymbol{\theta}^*\right)}, \tag{24}$$

where $\mathbf{x}_i = (1, x_{1i}, x_{2i}, ..., x_{Pi})$, $\mathbf{w}_i = (1, w_{1i}, w_{2i}, ..., w_{Qi})$ and $\mathbf{z}_i = (1, z_{1i}, z_{2i}, ..., z_{Ri})$ are vectors containing observations on, respectively, P, Q and R independent variables, and $\boldsymbol{\alpha}^* = (\alpha_0, \alpha_1, ..., \alpha_P)$, $\boldsymbol{\beta}^* = (\beta_0, \beta_1, ..., \beta_Q)$ and $\boldsymbol{\theta}^* = (\theta_0, \theta_1, ..., \theta_R)$ are vectors of unknown parameters. Observe that the covariate vectors $\mathbf{x}_i$, $\mathbf{w}_i$ and $\mathbf{z}_i$ could be exactly the same or different in each application.

2.3. Bayesian analysis

The Bayesian inference is an alternative to the maximum likelihood estimation of parameters. In the Bayesian approach it is necessary to specify a prior distribution for each unknown parameter (11). Applying the Bayes theorem, we have that the joint posterior density is given by the combination of the joint prior distribution and the likelihood function. Considering the proposed reparametrization of the model, we assume normal prior distributions for the parameters α_0, β_0 and θ_0. That is,

$$\ln\alpha \sim N(a_1, b_1), \tag{25}$$

$$\ln\beta \sim N(a_2, b_2) \tag{26}$$

and

$$\ln\frac{\theta}{1-\theta} \sim N(a_3, b_3), \tag{27}$$

where a_1, a_2, a_3, b_1, b_2 and b_3 are known hyperparameters, and $N(a,b)$ denotes a normal distribution with mean a and variance b.

In the case of the model in presence of covariates, consider the following prior distributions for the parameters: $\alpha_j \sim N(c_j, d_j)$, $\beta_k \sim N(e_k, f_k)$, and $\theta_l \sim N(g_l, h_l)$, where c_j, d_j, e_k, f_k, g_l and h_l $(j = 0,1,...,P,\quad k = 0,1,...,Q$ and $l = 0,1,...,R)$ are known hyperparameters.

Further, there is assumed prior independence among the parameters of the model. In this article, posterior summaries of interest were obtained using Markov chain Monte Carlo (MCMC) sampling available in the statistical software [14]. Details on the OpenBUGS code for fitting this model can be found on the Appendix at the end of the paper. Thus, 1,000,000 samples for each parameter of interest were generated, after a burn-in period of 5,000 iterations aimed to avoid the influence of the initial values and a thinning interval of 100 aimed to avoid correlation between successive samples. Convergence of the chains was assessed using visual examination of trace, density, and autocorrelation plots. Bayesian estimates of the parameters were obtained as the mean of the samples drawn from the joint posterior distribution and corresponding 95% credible intervals ($95\%CrI$) were given by the $0.025th$ and $0.975th$ percentiles of the posterior distributions.

2.4. Generating a random sample from a DDD

A sample of size n from a DDD with right-censored data can be randomly generated according to following steps:

- Fix values of α, β and θ.

- Generate n random samples from $M_i \sim Bernoulli(0,\theta)$.

- For $i=1,...,n$, consider $t_i' = \infty$ if $M_i = 0$ and $t_i' = F^{-1}(U_i;\boldsymbol{\lambda})$ if $M_i = 1$, where $F^{-1}(U_i;\boldsymbol{\lambda})$ is given by the expression (7) and $U_i \sim Uniform(0,\theta)$.

- Generate n random samples from $u_i' \sim Uniform(0,\max(t_i'))$, considering only the finite t_i.

- Let $t_i = \min(t_i', u_i')$. Pairs of simulated values (t_i, d_i), $i=1,...n$, are thus obtained, where $d_i = 1$ if $t_i < u_i'$ and $d_i = 0$ otherwise.

A similar algorithm was presented by Rocha et al. (17), considering other families of defective distributions. A R function based on these steps is presented in the Appendix at the end of the paper.

3. Results

3.1. A simulation study

A simulation study was carried out to assess the performance of the maximum likelihood estimation procedure for estimating the parameters of the DDD. It was estimated the coverage probability of the Wald-type confidence intervals for the parameters α, β and θ given by the expressions (19), (20) and (21), and the corresponding bias and mean squared errors (MSE). Considering the proposed reparametrization and the invariance property of the maximum likelihood estimators, MLE for α, β and θ are given by $\alpha_{ML} = \exp(\alpha_{0ML})$, $\beta_{ML} = \exp(\beta_{0ML})$ and $\theta_{ML} = \exp(\theta_{0ML})\left[1+\exp(\theta_{0ML})\right]^{-1}$. Coverage probability is defined as the proportion of times that the estimated confidence interval for a parameter of interest contains the true value of this parameter. If the confidence interval is adequate, we expect the coverage probability to be equal to nominal value. In this study the nominal confidence coefficient was 95%. The bias in the estimation of a parameter λ is estimated by $B^{-1}\sum_{b=1}^{B}\overset{(b)}{\lambda}_{ML} - \lambda_N$, where B is the number of simulated samples, $\overset{(b)}{\lambda}_{ML}$ is the maximum likelihood estimate for λ based on the b-th simulated sample ($b=1,...,B$), λ_N is the corresponding nominal value for λ and λ is a parameter that belongs to (α,β,θ). The corresponding MSE is estimated by $B^{-1}\sum_{b=1}^{B}(\overset{(b)}{\lambda}_{ML}-\lambda_N)^2$.

It was generated $B=5,000$ random samples each of size $n=25,30,35,\ 40,\ ...,\ 400$. Random samples were taken to come from (a) $T\sim DDD(2,2,0.8)$, (b) $T\sim DDD(1,3,0.5)$ and (c) $T\sim DDD(1.5,0.5,0.5)$. These choices were arbitrary. It was computed the maximum likelihood estimates $\overset{(b)}{\alpha}_{ML},\overset{(b)}{\beta}_{ML}$ and $\overset{(b)}{\theta}_{ML}$ and the respective standard errors for each simulated sample, and these quantities were used to estimate the coverage probability of the confidence intervals, the bias and the MSE.

Figure 2: Plots of the coverage probability, biases and MSE of $\overset{(b)}{\alpha}_{ML},\overset{(b)}{\beta}_{ML}$ and $\overset{(b)}{\theta}_{ML}$ versus n for simulated data from $T\sim DDD(2,2,0.8)$.

Figures 2 to 4 show the plots of the coverage probability, the biases and the MSE of $\overset{(b)}{\alpha}_{ML},\overset{(b)}{\beta}_{ML}$ and $\overset{(b)}{\theta}_{ML}$ versus n for simulated data from the DDD. We can observe that the MSE for all parameters generally decrease to zero with increasing n and the respective biases generally approach zero as n increases. For all parameters, the figures show that the MSE can assume very high values when $n<50$. The simulations show that the coverage probabilities for 95% confidence intervals for all parameters are satisfactory when the sample size n is larger than 100 , but not for small sample sizes.

Figure 3: Plots of the coverage probability, biases and MSE of $\overset{(b)}{\alpha}_{ML},\overset{(b)}{\beta}_{ML}$ and $\overset{(b)}{\theta}_{ML}$ versus n for simulated data from $T\sim DDD(1,3,0.5)$.

Figure 4: Plots of the coverage probability, biases and MSE of $\overset{(b)}{\alpha}_{ML},\overset{(b)}{\beta}_{ML}$ and $\overset{(b)}{\theta}_{ML}$ versus n for simulated data from $T\sim DDD(1.5,0.5,0.5)$.

Figure 5: Plots of the survival function $S(t;\boldsymbol{\lambda})$ estimated by the Kaplan-Meier method and by the frequentist and Bayesian parametric model based on the DDD. Simulated data. The horizontal dashed lines in each plot correspond to the nominal value for the cure fraction $(1-\theta)$ and vertical ticks are censored observations.

3.2. Applications to simulated data

We simulated samples of size $n=25$, 50 and 100 from the DDD for (a) (α,β,θ) = $(2,2,0.8)$, (b) (α,β,θ) = $(1,3,0.5)$ and (c) (α,β,θ) = $(1.5,0.5,0.5)$. Table 1 shows that the correspondent estimates for the parameters α, β and θ obtained by the maximum likelihood and Bayesian approaches are satisfactory close to each other. Figure 5 compares the plots of the survival function $S(t;\boldsymbol{\lambda})$ estimated by the Kaplan-Meier method and the curves obtained from the frequentist and Bayesian estimates (showed in Table 1) for the model based on DDD.

3.3. An application to real data

A study included 148 Brazilian women diagnosed and treated for invasive cervical carcinoma between 1992 and 2002 (3). In this application, it is assumed a subsample from this study, related to 118 women who received the standard treatment recommended by the International Federation of Gynecology and Obstetrics (FIGO). Let us consider as the outcome of interest the disease-free survival (DFS), defined as the time in complete months from the date of surgery to the first event of disease recurrence. Nearly 48% of the data are censored observations. Figure 6 shows a scatter plot of the observed disease-free survival times, where the censored data tend to be larger than the complete data, evidencing the presence of long-time survivors.

Considering these data, Table 2 shows maximum likelihood and Bayesian estimates for the parameters of the model based on DDD. In the Bayesian analysis, prior normal distributions $N(0,10)$ were assigned for $\ln\alpha$, $\ln\beta$ and $\ln\frac{\theta}{1-\theta}$. Note that maximum likelihood and Bayesian estimates are quite similar. Panel (a) of the Figure 7 shows the survival curves estimated by the Kaplan-Meier method and by the maximum likelihood and Bayesian approaches. In this graph, we note that the Kaplan-Meier curve seems to converge to a constant value greater than zero as the time t increases, thus suggesting the adequacy of a cure fraction model to these data.

Figure 6: Scatter plot of the observed disease-free survival times, considering censored and complete data.

In order to compare different defective probability distributions under a Bayesian approach, we calculate the log pseudo marginal likelihood (LPML) proposed by Gelfand et al. (10) as model selection technique. The larger the LPML value, the better the model fits the data. The LPML for the model based on the DDD distribution is -292.61. For the model based on the defective generalized Gompertz distribution (2; 15) we have $LPML = -295.45$, for the model based on the defective Gompertz distribution (12) we have $LPML = -295.27$ and for the model based on the defective inverse Gaussian distribution (1) we have $LPML = -290.19$. One can see that these values are close to each other, indicating that all these models seem to be equally adequate for the data.

Figure 7: Plots of the disease-free survival functions estimated from the Kaplan-Meier method and the frequentist and Bayesian models based on DDD, invasive cervical carcinoma data (3). Panel (a) shows the survival curves in the absence of covariates. Panel (b) shows the survival curves stratified by the clinical stage, where the parametric curves were obtaned from the model with covariates. Censored observations are marked by ticks on the survival curves.

In order to illustrate an application of the model based on the DDD in the presence of covariates, let us consider the clinical stage at the start of treatment, classified as I, II or III. This variable is included in the regression model by using two dummy variables, x_1 and x_2, where $x_1 = 0$ and $x_2 = 0$ if stage I, $x_1 = 1$ and $x_2 = 0$ if stage II, and $x_1 = 0$ and $x_2 = 1$ if stage III. Among the 118 women who had invasive cervical carcinoma, the disease was classified as clinical stage I in 39 (33%), 40 (34%) as clinical stage II, and 39 (33%) as clinical stage III.

Thus, the regression model for the data considers that

$$\alpha(\mathbf{x}) = \exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2), \tag{28}$$

$$\beta(\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \tag{29}$$

and

$$\theta(\mathbf{x}) = \frac{\exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}{1+\exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}. \tag{30}$$

Maximum likelihood and Bayesian estimates for the parameters α_0, α_1, α_2, β_0, β_1, β_2, θ_0, θ_1 and θ_2 of the regression model based on DDD are showed in Table 3. When comparing the results from both approaches, we can note that the estimates are a little different from each other, however, the panel (b) of the Figure 6 shows that the survival curves estimated from the maximum likelihood and Bayesian models are satisfactorily close to the empirical curves obtained from the Kaplan-Meier method. In addition, we also note in Table 3 that the credible intervals for β_2 and θ_2 do not include zero, evidencing a significant effect of the clinical stages on the disease-free survival times.

4. Conclusions

In the present article, we introduce a new defective distribution for survival analysis in the presence of a cure fraction. The great advantage of this distribution is that the cure fraction can be written as a function of only one parameter, unlike what happens with other usual defective distributions, such as the defective Gompertz distribution (12), the defective inverse Gaussian distribution (1) and the exponentiated-Weibull distribution (4). A simulation study showed that the maximum likelihood method has good statistical properties. Estimates of the parameters of the proposed distribution are not available in closed form, although they can be easily obtained through numerical iterative methods using the R software. To illustrate an application of the DDD, we used a real dataset from a cervical cancer study (3). We noted that the model satisfactorily fitted the data. As it was observed from the examples considering the simulated and real data, the model based on the new distribution is easy to be used in applications and it is a good alternative for the analysis of real time-to-event data in the presence of censored information and a cure fraction.

References

[1] Balka, J., Desmond, A. F., McNicholas, P. D., 2011. Bayesian and likelihood inference for cure rates based on defective inverse Gaussian regression models. *Journal of Applied Statistics*, 38, 127–144.

[2] Borges, P., 2017. EM algorithm-based likelihood estimation for a generalized Gompertz regression model in presence of survival data with long-term survivors: an application to uterine cervical cancer data. *Journal of Statistical Computation and Simulation*, 87, 1712–1722.

[3] Brenna, S. M., Silva, I. D., Zeferino, L. C., Pereira, J. S., Martinez, E. Z., Syrjänen, K. J., 2004. Prognostic value of P53 codon 72 polymorphism in invasive cervical cancer in Brazil. *Gynecologic Oncology*, 93, 374–380.

[4] Cancho, V. G., Bolfarine, H., 2001. Modeling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, 28, 659-671.

[5] Chen, M. H., Ibrahim, J. G., Sinha, D., 2002. Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis* 80, 101–126.

[6] Dagum, C., 1977. A new model of personal income distribution: specification and estimation. *Economie appliquée*, 30, 413–437.

[7] Domma, F., Latorre, G., Zenga, M., 2012. The Dagum distribution in reliability analisys. *Statistica & Applicazioni*, 10, 97–113.

[8] Domma, F., Giordano, S., Zenga, M., 2011. Maximum likelihood estimation in Dagum distribution with censored samples. *Journal of Applied Statistics*, 38, 2971–2985.

[9] Farewell, V. T., 1982. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38, 1041–1046.

[10] Gelfand, A. E., Dey, D. K., Chang, H., 1992. *Model determination using predictive distributions with implementation via sampling-based methods (with discussion)*, In Bayesian Statistics 4 (Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M., editors), Oxford University Press, 147-167.

[11] Gelman, A. Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., 2013. *Bayesian Data Analysis*, Third Edition, Chapman & Hall, Boca Raton.

[12] Gieser, P. W., Chang, M. N., Rao, P. V., Shuster, J. J., Pullen, J., 2014. Modelling cure rates using the Gompertz model with covariate information. *Statistics in Medicine*, 17, 831–839.

[13] Lambert, P. C., 2007. Modeling of the cure fraction in survival studies. *Stata*

Journal, 7, 1-25.

[14] Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

[15] Martinez, E. Z., Achcar, J. A., 2017. The defective generalized Gompertz distribution and its use in the analysis of lifetime data in presence of cure fraction, censored data and covariates. *Electronic Journal of Applied Statistical Analysis*, 10, 463–484.

[16] Rocha, R., Nadarajah, S., Tomazella, V., Louzada F., Eudes, A., 2015. New defective models based on the Kumaraswamy family of distributions with application to cancer data sets. *Statistical Methods in Medical Research*, 1–23.

[17] Rocha, R., Nadarajah, S., Tomazella, V., Louzada, F., 2017. A new class of defective models based on the Marshall-Olkin family of distributions for cure rate modeling. *Computational Statistics & Data Analysis*, 107, 48–63.

[18] Shoukri, M. M., Mian, I. U. M., Tracy, D. S., 1988. Sampling properties of estimators of the log-logistic distribution with application to Canadian precipitation data. *The Canadian Journal of Statistics*, 16, 223–236.

[19] Xiang, L., Ma, X., Yau, K. K., 2011. Mixture cure model with random effects for clustered interval-censored survival data. *Statistics in Medicine*, 30, 995–1006.

[20] Wienke, A., Lichtenstein, P., Yashin, A. I., 2003. A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. *Biometrics*, 59, 1178–1183.

Appendix

The following R function rDDD can be used to generate random samples of size n from a DDD with parameters α , β and θ .

```
 rDDD <- function(n,alpha,beta,theta) {
m <- rbinom(n,prob=theta,size=1)
u <- runif(n,0,theta)
y0 <- (u*theta/(beta*(theta-u)))^(1/alpha)
t0<-ifelse(m,y0,Inf)
maxti<-max(y0*m)
w <- runif(n,0,maxti)
t <- pmin(t0,w)
d <- as.numeric(t0<w)
data <- data.frame(t,d)
return (data) }
```

The following R code uses the function maxLik of the maxLik library to find the maximum likelihood estimates of the DDD. In this R code, St is the survival function, ft is the probability density function, like is the likelihood function, t is the time to event and d is the censoring indicator. In addition, limlow and limupp are the lower and upper limits of the 95% confidence intervals for the parameters of the model.

```
log.f <- function(parms){
alpha0 <- parms[1]
beta0 <- parms[2]
theta0 <- parms[3]
alpha <- exp(alpha0)
beta <- exp(beta0)
theta <- exp(theta0)/(1+exp(theta0))
St <- 1-theta*beta/(beta+theta*t^(-alpha))
ft <- alpha*beta*theta*theta
      *t^(-(alpha+1))/((beta+theta*t^(-alpha))^2)
like <- ft^d * St^(1-d)
L <- sum(log(like))
if (is.na(L)==TRUE) {return(-Inf)} else {return(L)}
}
library(maxLik)
mle <- maxLik(logLik=log.f,start=c(0.5,0.5,0.5))
summary(mle)
S<-vcov(mle)
limlow<-limupp<-rep(NA,3)
limlow[1] <- exp(mle$estimate[1] - qnorm(0.975) * sqrt(S[1,1]))
limupp[1] <- exp(mle$estimate[1] + qnorm(0.975) * sqrt(S[1,1]))
limlow[2] <- exp(mle$estimate[2] - qnorm(0.975) * sqrt(S[2,2]))
limupp[2] <- exp(mle$estimate[2] + qnorm(0.975) * sqrt(S[2,2]))
```

```
        limlow[3] <- 1-1/(1+exp(mle$estimate[3] - qnorm(0.975) * sqrt(S[3,3])))
        limupp[3] <-
        1-1/(1+exp(mle$estimate[3] + qnorm(0.975) * sqrt(S[3,3])))
est<-c(exp(mle$estimate[1]),exp(mle$estimate[2]),
        exp(mle$estimate[3])/(1+exp(mle$estimate[3])))
cbind(est,limlow,limupp)
```

The OpenBUGS software only requires the specification of the distribution for the data and the prior distribution for the parameters. The following OpenBUGS code can be used to obtain the Bayesian estimates for the DDD. In this OpenBUGS code, N denotes the sample size, S[i] is the survival function and L[i] is the likelihood function.

```
         model {
        for (i in 1:N) {
        S[i] <- 1 - theta*beta/(beta+theta*pow(t[i],-alpha))
        f[i] <- alpha*beta*theta*theta*pow(t[i],-(alpha+1))/
        (pow(beta+theta*pow(t[i],-alpha),2))
        L[i] <- pow(f[i],d[i])*pow(S[i],1-d[i])
        logL[i] <- log(L[i])
        zeros[i] <- 0
        zeros[i] ~ dloglik(logL[i]) }
        log(alpha) <- a0
        log(beta) <- b0
        logit(theta) <- th0
        a0 ~ dnorm(0,prec)
        b0 ~ dnorm(0,prec)
        th0 ~ dnorm(0,prec)
        }
        # Inits
        list(a0=1,b0=1,th0=0.5)
```

All the presented codes can be readily adapted to include covariates. The data from the cervical carcinoma study in OpenBUGS format are shown below, including the dummy variables X1 and X2 used in the model. This dataset can be freely used by students and researchers on their own works, but the article from Brenna et al. (3) should be cited as its source.

```
 # Data
list(N=118,t=c
(28, 8, 116, 2, 1, 112, 7, 17, 32, 1, 19, 5, 29, 42, 91, 8, 91, 85, 2, 1, 89, 3, 1, 5, 1, 84, 1, 112, 99,
1, 6, 2, 62, 6, 103, 24, 1, 4, 19, 24, 26, 1, 8, 2, 2, 1, 77, 27, 65, 2, 64, 8, 16, 91, 54, 90, 60, 8, 1, 1,
93, 3, 101, 10, 5, 12, 1, 85, 37, 5, 89, 2, 91, 88, 44, 19, 1, 15, 69, 88, 25, 36, 86, 45, 7, 22, 90, 2,
49, 9, 8, 66, 6, 15, 11, 1, 46, 23, 5, 28, 4, 4, 15, 13, 1, 1, 6, 32, 49, 23, 2, 39, 19, 37, 55, 43, 1, 1),
d=c (0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1,
0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1,
0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1),
x1=c (1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0,
0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0,
0, 0, 0, 1),
x2=c (0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1,
1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1,
0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0))
```

FIGURE1

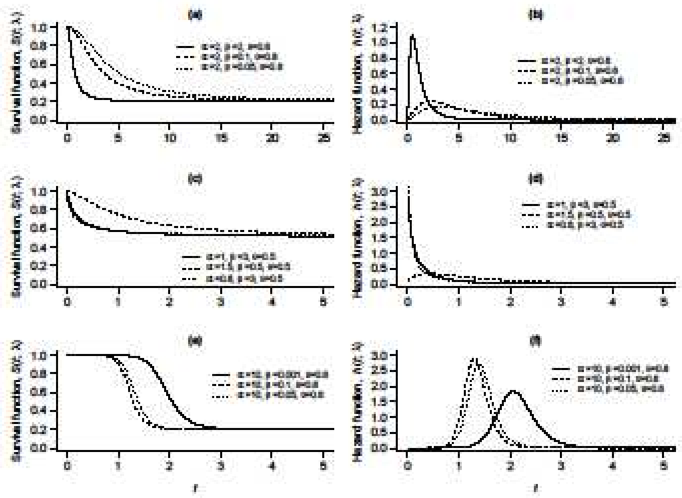

Figure 1: Plots of the survival (left panels) and hazard functions (right panels) for the DDD, based on some choices for the parameters α, β and θ.

FIGURE2

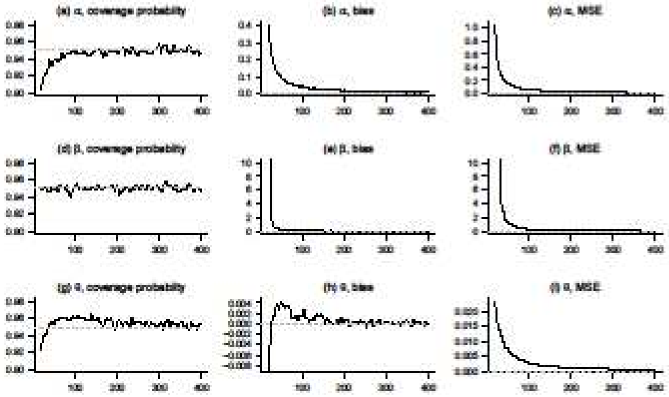

Figure 2: Plots of the coverage probability, biases and MSE of $\widehat{\alpha}_{ML}^{(b)}$, $\widehat{\beta}_{ML}^{(b)}$ and $\widehat{\theta}_{ML}^{(b)}$ versus n for simulated data from $T \sim DDD(2, 2, 0.8)$.

FIGURE3

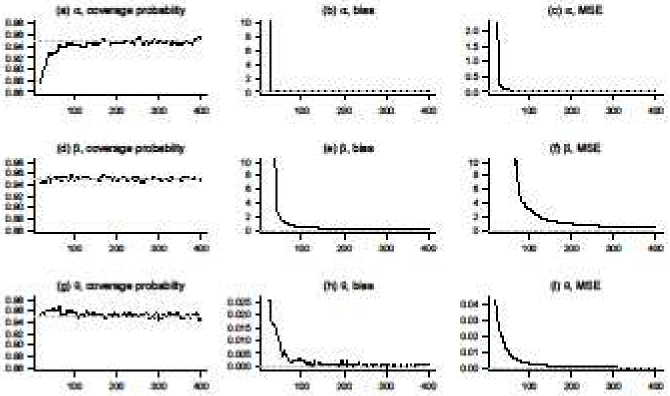

Figure 3: Plots of the coverage probability, biases and MSE of $\widehat{\alpha}_{ML}^{(b)}$, $\widehat{\beta}_{ML}^{(b)}$ and $\widehat{\theta}_{ML}^{(b)}$ versus n for simulated data from $T \sim DDD(1, 3, 0.5)$.

FIGURE4

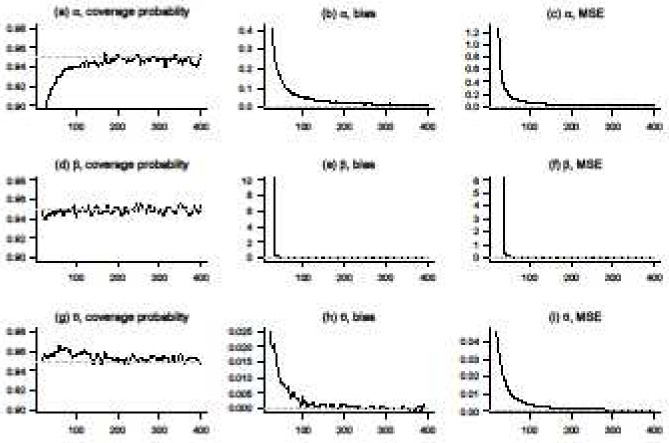

Figure 4: Plots of the coverage probability, biases and MSE of $\widehat{\alpha}_{ML}^{(b)}$, $\widehat{\beta}_{ML}^{(b)}$ and $\widehat{\theta}_{ML}^{(b)}$ versus n for simulated data from $T \sim DDD(1.5, 0.5, 0.5)$.

FIGURE5

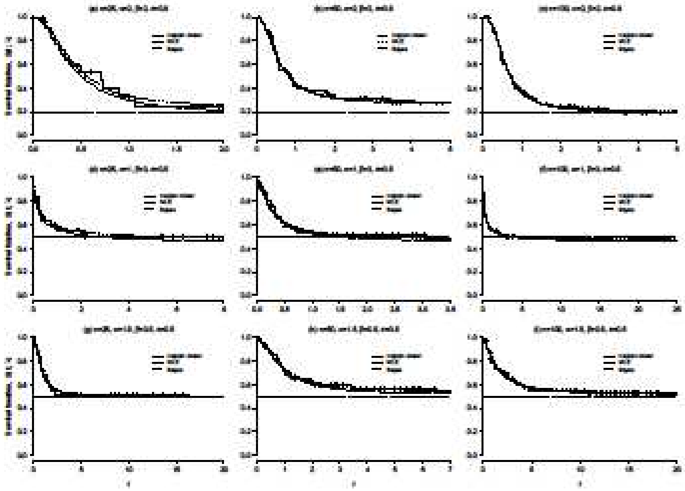

Figure 5: Plots of the survival function $S(t;\lambda)$ estimated by the Kaplan-Meier method and by the frequentist and Bayesian parametric model based on the DDD. Simulated data. The horizontal dashed lines in each plot correspond to the nominal value for the cure fraction $(1-\theta)$ and vertical ticks are censored observations.

FIGURE6

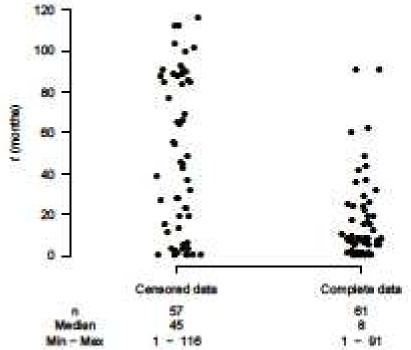

Figure 6: Scatter plot of the observed disease-free survival times, considering censored and complete data.

FIGURE7

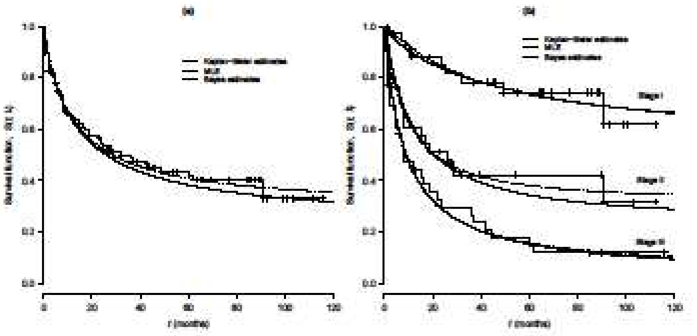

Figure 7: Plots of the disease-free survival functions estimated from the Kaplan-Meier method and the frequentist and Bayesian models based on DDD, invasive cervical carcinoma data (3). Panel (a) shows the survival curves in the absence of covariates. Panel (b) shows the survival curves stratified by the clinical stage, where the parametric curves were obtaned from the model with covariates. Censored observations are marked by ticks on the survival curves.

Table 1: Maximum likelihood and Bayesian estimates for the parameters of the DDD. Simulated data.

		Nominal	Maximum likelihood estimates			Bayesian estimates		
n	Parameter	values	Estimate	*95%CI*	*AIC*	Estimate	*95%CrI*	*DIC*
25	α	2	2.1053	(1.2764 , 3.4724)	30.3	1.9200	(1.0630 , 3.0030)	29.5
	β	2	4.8960	(1.0888 , 22.01)		4.7570	(0.9937 , 16.7)	
	θ	0.8	0.7711	(0.4449 , 0.9340)		0.8249	(0.5499 , 0.9978)	
50	α	2	2.1713	(1.5501 , 3.0414)	94.6	2.1130	(1.3810 , 2.8880)	94.8
	β	2	2.4187	(1.0966 , 5.3343)		2.4430	(0.9963 , 5.0690)	
	θ	0.8	0.7231	(0.5606 , 0.8423)		0.7337	(0.5796 , 0.8886)	
100	α	2	2.5041	(2.0140 , 3.1133)	157.6	2.4780	(1.9580 , 3.0450)	157.7
	β	2	3.1553	(1.8236 , 5.4592)		3.1880	(1.7700 , 5.4030)	
	θ	0.8	0.8010	(0.6927 , 0.8778)		0.8035	(0.7062 , 0.8912)	
25	α	1	0.8389	(0.4607 , 1.5274)	41.6	0.8022	(0.3982 , 1.3370)	40.9
	β	3	1.7689	(0.4240 , 7.3792)		1.9580	(0.4172 , 6.4700)	
	θ	0.5	0.5381	(0.3047 , 0.7559)		0.5666	(0.3442 , 0.8223)	
50	α	1	1.4443	(0.9342 , 2.2327)	59.9	1.3810	(0.8084 , 2.0460)	59.7
	β	3	3.4745	(0.9990 , 12.0831)		3.6430	(0.8813 , 10.9800)	
	θ	0.5	0.5279	(0.3574 , 0.6921)		0.5481	(0.3770 , 0.7546)	
100	α	1	0.8810	(0.6484 , 1.1135)	161.6	0.8713	(0.6443 , 1.1140)	161.7
	β	3	2.3120	(0.7226 , 3.9013)		2.3650	(1.1220 , 4.4650)	
	θ	0.5	0.5362	(0.4317 , 0.6406)		0.5404	(0.4354 , 0.6481)	
25	α	1.5	1.8627	(1.1255 , 3.0825)	63.4	1.8250	(0.9496 , 2.8740)	63.3

	β	0.5	0.7607	(0.2594 , 2.2305)		0.8588	(0.2497 , 2.2280)	
	θ	0.5	0.4960	(0.3049 , 0.6882)		0.5039	(0.3171 , 0.6948)	
50	α	1.5	1.7934	(1.1371 , 2.8279)	99.2	1.7300	(0.9563 , 2.6140)	99.1
	β	0.5	0.6306	(0.2625 , 1.5148)		0.6562	(0.2423 , 1.4540)	
	θ	0.5	0.4598	(0.3013 , 0.6268)		0.4862	(0.3240 , 0.6992)	
100	α	1.5	1.5017	(1.1101 , 2.0314)	256.9	1.4770	(1.0250 , 1.9580)	256.9
	β	0.5	0.3137	(0.1789 , 0.5497)		0.3253	(0.1757 , 0.5450)	
	θ	0.5	0.4778	(0.3666 , 0.5913)		0.4879	(0.3793 , 0.6060)	

Table 2: Maximum likelihood and Bayesian estimates for the parameters of the DDD. Invasive cervical carcinoma data (3).

	Maximum likelihood estimates			Bayesian estimates		
Parameter	Estimate	*95%CI*	*AIC*	Estimate	*95%CrI*	*DIC*
α	0.9825	(0.7407 , 1.3033)	585.8	0.9218	(0.6591 , 1.2230)	585.2
β	0.0619	(0.0326 , 0.1172)		0.0698	(0.0342 , 0.1206)	
θ	0.7126	(0.5316 , 0.8441)		0.7735	(0.5938 , 0.9943)	

Table 3: Maximum likelihood and Bayesian estimates for the parameters of the regression model based on DDD. Invasive cervical carcinoma data (3).

	Maximum likelihood estimates			Bayesian estimates		
Parameter	Estimate	*95%CI*	*AIC*	Estimate	*95%CrI*	*DIC*
α_0	0.1236	(-0.7552 , 1.0025)	566.1	-0.1591	(-0.7795 , 0.4027)	562.0
α_1	0.0057	(-0.9620 , 0.9733)		0.2162	(-0.4802 , 0.9468)	
α_2	-0.1381	(-1.1149 , 0.8388)		0.1400	(-0.4926 , 0.8103)	
β_0	-5.0141	(-7.7182 , -2.3099)		-4.1410	(-5.9450 , -2.6000)	
β_1	2.1613	(-0.6714 , 4.9941)		1.2880	(-0.5481 , 3.2950)	
β_2	2.9022	(0.0691 , 5.7352)		1.9900	(0.2676 , 3.9040)	
θ_0	-0.2806	(-2.3972 , 1.8360)		0.1008	(-0.9615 , 1.4970)	
θ_1	1.0744	(-1.2878 , 3.4366)		1.1130	(-0.3124 , 2.6970)	
θ_2	3.4974	(-2.1249 , 9.1197)		3.5840	(1.9220 , 5.4140)	