Austrian Journal of Statistics
January 2022, Volume 51, 144–168.
http://www.ajs.or.at/
doi:10.17713/ajs.v51i2.1285

The Bivariate Defective Gompertz Distribution Based on Clayton Copula with Applications to Medical Data

Marcos Vinicius de Oliveira Peres
University of São Paulo

Ricardo Puziol de Oliveira
State University of Maringá

Jorge Alberto Achcar
University of São Paulo

Edson Zangiacomi Martinez
University of São Paulo

Abstract

In medical studies, it is common the presence of a fraction of patients who do not experience the event of interest. These patients are people who are not at risk of the event or are patients who were cured during the research. The proportion of immune or cured patients is known in the literature as cure rate. In general, the traditional existing lifetime statistical models are not appropriate to model data sets with cure rate, including bivariate lifetimes. In this paper, it is proposed a bivariate model based on a defective Gompertz distribution and also using a Clayton copula function to capture the possible dependence structure between the lifetimes. An extensive simulation study was carried out in order to evaluate the biases and the mean squared errors for the maximum likelihood estimators of the parameters associated to the proposed distribution. Some applications using medical data are presented to show the usefulness of the proposed model. Maximum likelihood and Bayesian methods were used to estimate the parameters of the model.

Keywords: Clayton copula, cure rate, defective Gompertz distribution, survival analysis.

1. Introduction

The use of survival statistical models for time-to-event data is common in several areas of study, especially in medical research. Traditional parametric and non-parametric tools, such as, Kaplan-Meier estimator for the survival function, log-rank and Wilcoxon tests and the semi-parametric Cox proportional hazard model, are widely used in medical data analysis (see, e.g., Kleinbaum and Klein 2012). These methods assume that all individuals are susceptible to the event of interest. However, for example in clinical studies, there may be patients who will not experience the event under investigation, that is, these patients are immune to the event or they were cured during the research. This situation is suggested when a Kaplan-Meier estimator plot for the survival function describes a behavior with stable plateau and large censored data at the right of the curve (Corbière, Commenges, Taylor, and Joly 2009; Wienke 2010). In this way, the use of models that incorporate this plateau, named cure rate models, could be a better alternative to predict or to identify prognostics factors that affects the survival probability.

According to Vahidpour (2016), there are at least two kinds of models for data with cure fraction: the mixture cure rate models, also known as standard cure rate models (see, for example, De Angelis, Capocaccia, Hakulinen, Soderman, and Verdecchia 1999; Tsodikov, Ibrahim, and Yakovlev 2003; Lambert, Thompson, Weston, and Dickman 2006), and the non-mixture cure rate models, which are not so popular (see Achcar, Coelho-Barros, and Mazucheli 2012; Vahidpour 2016). Let us denote by T, the time for the occurrence of the event of interest. Following Maller and Zhou (1996), the standard cure rate model assuming that the probability of the time-to-event to be greater than a specified time t is given by the survival function,

$$S(t) = \rho + (1 - \rho)S_0(t) \tag{1}$$

where $\rho \in (0, 1)$ is the mixing parameter which represents the proportion of "long-term survivors", "non-susceptible" or "cured patients", and $S_0(t)$ denotes a proper survival function for the non-cured or susceptible group in the population. Observe that if $t \to \infty$, then $S(t) \to \rho$, that is, the survival function has an asymptote at the cure rate ρ.

On other hand, the non-mixture model defines an asymptote for the survival function, that is associated to the cure rate (see, Tsodikov *et al.* 2003). In this case, the survival function for the non-mixture cure rate model is given by,

$$S(t) = \rho^{F_0(t)} = \exp\{\ln(\rho)F_0(t)\} \tag{2}$$

where $\rho \in (0, 1)$ is the probability of cured patients and $F_0(t) = 1 - S_0(t)$ denotes a proper distribution function for the non-cured or susceptible group in the population.

Different approaches have been presented in the literature to model cure rate, especially for univariate lifetime data: Boag (1949), Ghitany and Maller (1992), De Angelis *et al.* (1999), Chen, Ibrahim, and Sinha (2002), Lambert *et al.* (2006), Castro, Cancho, and Rodrigues (2009), Chen, Ibrahim, and Sinha (1999), Achcar *et al.* (2012) and Martinez, Achcar, Jácome, and Santos (2013). However, the cure rate models are not the only ones to deal with long-term survivors, we also could use, as an alternative, the defective models. The main property of a proper probability distribution is that $\lim_{t\to\infty} F(t) = 1$ and, consequently, $\lim_{t\to\infty} S(t) = 0$. For defective models, the survival function, $S(t)$, converges to a value ρ, where ρ denotes the cure rate. Some approaches for defective models can been found in: Cancho and Bolfarine (2001), Balka, Desmond, and McNicholas (2011), da Rocha, Tomazella, and Louzada (2014), dos Santos, Achcar, and Martinez (2017), Rocha, Nadarajah, Tomazella, and Louzada (2017a), Martinez and Achcar (2018), among others.

In some studies, the main objective may be related to analyze the lifetime data assuming two time-to-event variables. As a special situation, we could be interested in the times of occurrence of a specified event, that could be reinfection, in the treatment of both lungs where we could use univariate lifetime models assuming independence between both time-to-event variables. However, in this situation, the times could not be independent since the patient needs both lungs working to survive. In this case, there may be the presence of a dependence structure that is not present when using univariate analyzes associated with each response, which is a motivation for the use of bivariate models. Different bivariate parametric models are introduced in the literature for the analysis of bivariate lifetime data: Marshall and Olkin (1967), Block and Basu (1974), Vaupel, Manton, and Stallard (1979) and Block and Basu (1974), Wienke, Lichtenstein, and Yashin (2003), Yu and Peng (2008), Achcar, Coelho-Barros, and Mazucheli (2013), Fachini, Ortega, and Cordeiro (2014) and de Oliveira, Achcar, Peralta, and Mazucheli (2019). As an alternative, the dependence structure can be specified by using a Copula function due to its simplicity. According to Hofert, Kojadinovic, Mächler, and Yan (2019) a copula is a multivariate distribution function with standard uniform univariate marginals. Many copula functions are considered to model data with cure rate: Wienke, Locatelli, and Yashin (2006), Li, Tiwari, and Guha (2007), Fachini *et al.* (2014), Martinez and Achcar (2014), Coelho-Barros, Achcar, and Mazucheli (2016) and Achcar, Martinez, and Tovar Cuevas (2016). More recently, Peres, Achcar, and Martinez (2020) conducted a comprehensive review of fifteen different copula functions that can be used to model survival data.

The main goal of this paper is to explore the use of the Clayton copula in the analysis of bivariate lifetime data assuming a bivariate defective Gompertz distribution to estimate the cure rate. Different correlation values between the time-to-event variables are considered in a simulation study that was done in order to describe the behavior of the dependence structure of the proposed model. The maximum likelihood method using existing numerical optimization algorithms was considered to get the inferences of interest under a frequentist approach and MCMC (Markov Chain Monte Carlo) simulation methods, as the popular Gibbs sampling and Metropolis-Hastings algorithms, were used to get the posterior summaries of interest under a Bayesian approach (Gelfand and Smith 1990; Chib and Greenberg 1995). The paper is organized as follows: in Section 2, it is presented the proposed methodology using the Clayton copula as well the inference methods. The simulation procedures and the obtained results are showed in Section 3. In Section 4, four applications related to real medica data are presented, using the proposed methodology. Finally, Section 5 closes the paper with some concluding remarks.

2. Statistical methods

2.1. Univariate defective Gompertz distribution

The main property of defective models is a survival function $S(t)$ that converges to a value ρ as t tends to infinity, where ρ denotes the cure rate parameter. Cantor and Shuster (1992) introduced the two-parameter defective Gompertz (DG) distribution also studied by Gieser, Chang, Rao, Shuster, and Pullen (1998) and dos Santos *et al.* (2017). The survival function for the DG distribution is given by,

$$S(t) = \exp\left\{-\frac{\alpha}{\beta}\left[1 - \exp(-\beta t)\right]\right\}, \tag{3}$$

where $t > 0$ and $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter. Taking the limit of the survival function from the DG distribution, the cure rate parameter ρ is given by

$$\rho = \lim_{t\to\infty} S(t) = \exp\left\{-\frac{\alpha}{\beta}\right\}. \tag{4}$$

The correspondent probability density and hazard function are respectively given by

$$f(t) = \alpha \exp(-\beta t) \exp\left\{-\frac{\alpha}{\beta}\left[1 - \exp(-\beta t)\right]\right\} \quad \text{and} \quad h(t) = \alpha \exp(-\beta t). \tag{5}$$

Note that the hazard function has only decreased shape, and this is an important limitation of a model based on the DG distribution.

2.2. Copula functions

Copula functions are used to create a joint distribution function of two or more marginal univariate distributions following standard uniform distribution U$(0, 1)$ to form a multivariate distribution (Nelsen 2007). Considering a m-variate function F, the respective copula is a function $C : [0, 1]^m \to [0, 1]$ that satisfies

$$F(y_1, \ldots, y_m) = C(F_1(y_1), ..., F_m(y_m); \phi) = C_\phi(F_1(y_1), ..., F_m(y_m)), \tag{6}$$

where ϕ is a parameter that measures the dependence between the marginals. The join probability density is given by

$$f(y_1, \dots, y_m) = c_\phi(F_1(y_1), ..., F_m(y_m)) \prod_{i=1}^{m} f_i(y_i), \tag{7}$$

where $f_i(y_i),\ i = 1, ..., m$, are the marginal density functions and $c_\phi(F_1(y_1), ..., F_m(y_m))$ is the derivative of order m of (6) in relation to $y_1, ..., y_m$. If the random variables are independent, then $c_\phi(F_1(y_1), ..., F_m(y_m)) = 1$.

For the bivariate case ($m = 2$) and under the context of survival analysis, considering $S_1(t_1)$ and $S_2(t_2)$ as the univariate survival functions, the bivariate joint survival function $S(t_1, t_2)$ is defined by a copula function given by

$$S(t_1, t_2) = C_\phi(S_1(t_1), S_2(t_2)), \tag{8}$$

for $t_1 > 0$ and $t_2 > 0$, with the respective joint probability density function given by

$$f(t_1, t_2) = \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} = f_1(t_1) f_2(t_2) c_\phi(S_1(t_1), S_2(t_2)), \tag{9}$$

where $c_\phi(u, v)$ is the copula density function defined by

$$c_\phi(u, v) = \frac{\partial^2}{\partial u \partial v} C_\phi(u, v), \tag{10}$$

where $u = S_1(t_1)$ and $v = S_2(t_2)$.

The estimation of the correlation between two random variables using copula functions usually is made using the Kendall's tau (τ_k) and Spearman's rho (τ_s). According to Joe (2014), those coefficients can be expressed by the equations

$$\tau_k = 1 - 4 \int_0^1 \int_0^1 \frac{\partial C_\phi(u, v)}{\partial u} \frac{\partial C_\phi(u, v)}{\partial v} \, \mathrm{d}u\mathrm{d}v \tag{11}$$

and

$$\tau_s = 12 \int_0^1 \int_0^1 C_\phi(u, v) \, \mathrm{d}u\mathrm{d}v - 3 \tag{12}$$

The literature introduces many copula functions which could be considered to build different bivariate lifetime distributions. However, it is important to choose copula functions suitable for each type of dependence structure in the applications. In each application, it is possible to obtain some information on the dependence structure by an exploratory graphical analysis, but unfortunately this can be difficult in some cases. Another framework that can help in choosing the copula function is to determine the empirical correlation between the random variables. This correlation can be obtained through iterative multiple imputation (Schemper, Kaider, Wakounig, and Heinze 2013). In the present study, we explore the Clayton copula function as a special case when appropriate in the data analysis. The Clayton copula is a popular choice to be fitted by bivariate time-to-event data, due its ability to describe positive dependence.

The Clayton copula was first introduced by Clayton (1978) and later studied by Cook and Johnson (1981) and Oakes (1982). Assuming this copula function, the joint survival function $S(t_1, t_2)$ is given by

$$S(t_1, t_2) = \left\{ [S_1(t_1)]^{-\phi} + [S_2(t_2)]^{-\phi} - 1 \right\}^{-1/\phi}, \tag{13}$$

where $S_1(t_1)$ and $S_2(t_2)$ are, respectively, the marginal survival functions for the random variables T_1 and T_2 and $\phi \in (0, \infty)$. When $\phi \to 0$, there is an indication that T_1 and T_2 are independent. The relationship between the copula parameter ρ and the dependence structure can be interpreted by the Spearman's correlation $\tau_s(\phi)$. However, obtaining this measure using the equation (12) can be a difficult task. The Kendall's correlation coefficient is given by

$$\tau_k(\phi) = \frac{\phi}{\phi + 2}. \tag{14}$$

Note that $0 < \tau(\phi) \leq 1$, where if $\phi \to \infty$, we have total dependence between T_1 and T_2. The Clayton copula, is thus adequate to model positive dependences and it has the advantage of measuring a wide range of positive correlations. The respective joint probability density function for T_1 and T_2 is given by

$$f(t_1, t_2) = f_1(t_1) f_2(t_2)(1+\phi) \left[S_1(t_1)S_2(t_2)\right]^{-1-\phi} \left\{[S_1(t_1)]^{-\phi} + [S_2(t_2)]^{-\phi} - 1\right\}^{-2-1/\phi}, \tag{15}$$

where $f_1(t_1)$ and $f_2(t_2)$ are, respectively, the marginal probability density functions for the random variables T_1 and T_2.

2.3. Bivariate defective Gompertz distribution

The marginal probability density and survival functions for the lifetimes T_j $(j = 1, 2)$ considering the DG distribution are given, respectively, by

$$f_j(t_j) = \alpha_j \exp(-\beta_j t_j) \exp\left\{-\frac{\alpha_j}{\beta_j}\left[1 - \exp(-\beta_j t_j)\right]\right\} \tag{16}$$

and

$$S_j(t_j) = \exp\left\{-\frac{\alpha_j}{\beta_j}\left[1 - \exp(-\beta_j t_j)\right]\right\}. \tag{17}$$

Thus, the correspondent cure rates are given by

$$\rho_j = \exp\left\{-\frac{\alpha_j}{\beta_j}\right\}, \tag{18}$$

where j is equal to 1 or 2, corresponding to the time-to-event variables T_1 and T_2, respectively.

The joint survival and density functions for the bivariate defective Gompertz distribution using a Clayton copula function (13) (BDGD) are given, respectively, by

$$S(t_1, t_2) = \left\{\left[\exp\left\{-\frac{\alpha_1}{\beta_1}\left[1 - \exp(-\beta_1 t_1)\right]\right\}\right]^{-\phi} + \left[\exp\left\{-\frac{\alpha_2}{\beta_2}\left[1 - \exp(-\beta_2 t_2)\right]\right\}\right]^{-\phi} - 1\right\}^{-1/\phi}, \tag{19}$$

and,

$$\begin{aligned} f(t_1, t_2) \;&=\; \alpha_1\alpha_2 \exp(-\beta_1 t_1 - \beta_2 t_2) \exp\left\{-\frac{\alpha_1}{\beta_1}\left[1 - \exp(-\beta_1 t_1)\right] - \frac{\alpha_2}{\beta_2}\left[1 - \exp(-\beta_2 t_2)\right]\right\} \\ &\times\; (1+\phi)\left[\exp\left\{-\frac{\alpha_1}{\beta_1}\left[1 - \exp(-\beta_1 t_1)\right] - \frac{\alpha_2}{\beta_2}\left[1 - \exp(-\beta_2 t_2)\right]\right\}\right]^{-1-\phi} \\ &\left\{\left[\exp\left\{-\frac{\alpha_1}{\beta_1}\left[1 - \exp(-\beta_1 t_1)\right]\right\}\right]^{-\phi} + \left[\exp\left\{-\frac{\alpha_2}{\beta_2}\left[1 - \exp(-\beta_2 t_2)\right]\right\}\right]^{-\phi} - 1\right\}^{-2-1/\phi} \end{aligned} \tag{20}$$

2.4. Inference methods

Maximum likelihood estimation

To obtain the bivariate likelihood function, let us assume a random sample of size n, where each sample has two lifetimes T_1 and T_2. Let us consider that both T_1 and T_2 can be right-censored and that this censoring is independent of each time-to-event. For each i^{th} observation $(i = 1, \ldots, n)$ it is possible to classify the data into one of four classes given by,

(1) $C1$: both t_{1i} and t_{2i} are uncensored lifetimes;

(2) $C2$: t_{1i} is a complete lifetime and t_{2i} is a censored lifetime;

(3) $C3$: t_{2i} is a complete lifetime and t_{1i} is a censored lifetime;

(4) $C4$: t_{1i} and t_{2i} are censored lifetimes.

Thus, the likelihood function is given by

$$L = \prod_{i \in C_1} [f(t_{1i}, t_{2i})] \prod_{i \in C_2} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right] \prod_{i \in C_3} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right] \prod_{i \in C_4} [S(t_{1i}, t_{2i})], \qquad (21)$$

where $f(t_1, t_2)$ is the joint probability function of T_1 and T_2, given in equation (15) and $S(t_1, t_2)$ is the joint survival function given by equation (13) considering the Clayton copula. Let us consider two indicator variables, denoted by δ_{1i} and δ_{2i}, where $\delta_{ki} = 1$ when t_{ki} is an observed lifetime and $\delta_{ki} = 0$ when t_{ki} a censored observation, $k = 1, 2$ and $i = 1, ..., n$. In this way, it is possible to rewrite the likelihood function as

$$L = \prod_{i=1}^{n} [f(t_{1i}, t_{2i})]^{\delta_{1i}\delta_{2i}} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right]^{\delta_{1i}(1-\delta_{2i})} \left[-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right]^{\delta_{2i}(1-\delta_{1i})} [S(t_{1i}, t_{2i})]^{(1-\delta_{1i})(1-\delta_{2i})} . (22)$$

In the absence of censored observations, the expression above is reduced to the form,

$$L = \prod_{i=1}^{n} \frac{\partial^2 S(t_{1i}, t_{2i})}{\partial t_{1i} \partial t_{2i}} = \prod_{i=1}^{n} f(t_1, t_2). \qquad (23)$$

For the Clayton copula, the first partial derivatives of $S(t_1, t_2)$ with respect to t_1 and t_2 are given by the following relations,

$$-\frac{\partial S(t_1, t_2)}{\partial t_1} = f_1(t_1) S_1(t_1)^{-(\phi+1)} \left[S_1(t_1)^{-\phi} + S_2(t_2)^{-\phi} - 1 \right]^{-(1+1/\phi)} \qquad (24)$$

and

$$-\frac{\partial S(t_1, t_2)}{\partial t_2} = f_2(t_2) S_2(t_2)^{-(\phi+1)} \left[S_1(t_1)^{-\phi} + S_2(t_2)^{-\phi} - 1 \right]^{-(1+1/\phi)} . \qquad (25)$$

Bayesian analysis

Assuming the proposed model, let $\boldsymbol{\theta} = (\alpha_1, \beta_1, \alpha_2, \beta_2, \phi)$ be the vector of unknown parameters. Under a Bayesian framework, the joint posterior distribution for the model parameters is obtained by combining the joint prior distribution of the parameters and the likelihood function given by equation (22) (Gelman, Stern, Carlin, Dunson, Vehtari, and Rubin 2013). To simulate samples from the joint posterior distribution, we could consider the use of MCMC (Markov Chain Monte Carlo) algorithms implemented in the R2jags package (Plummer *et al.* 2003) in *R* software, where we just need to specify the data distribution and the prior distribution for the parameters.

Under a Bayesian approach, we assume independent uniform prior distributions for the parameters α_1, β_1 α_2, β_2 and ϕ. That is, we assume $\alpha_1 \sim Unif(a_1, b_1)$, $\alpha_2 \sim Unif(a_2, b_2)$, $\beta_1 \sim Unif(a_3, b_3)$, $\beta_2 \sim Unif(a_4, b_4)$ and $\phi \sim Unif(a_6, b_6)$, where a_k and b_k,$k = 1, ..., 4$, are known hyperparameters, and $Unif(a, b)$ denotes a uniform distribution with mean $(a + b)/2$ and variance $(a + b)^2/12$. The values of hyperparameters a and b were chosen in order to reflect prior knowledge of experts and better performance of the MCMC algorithm in terms of good convergence. These values were obtained using empirical Bayesian methods (Carlin and Louis 2000) as information on the cure rate obtained from the non-parametrical Kaplan-Meier estimator for the survival function and information on the correlation obtained from empirical estimators.

3. Simulation study

The simulation study was carried out in order to evaluate the performance of the maximum likelihood (ML) estimation. The coverage probability of the Wald confidence intervals for the parameters $\alpha_1,\ \alpha_2,\ \beta_1,\ \beta_2,\ \rho_1,\ \rho_2$ and ϕ, with their corresponding bias and mean squared errors (MSE) were considered. Calculations of the coverage probabilities were carried out for a nominal coverage of 95%, corresponding to 95 successes in each 100 simulated samples. Since ρ_1 and ρ_2 are functions of other parameters, the Wald confidence interval for these parameters were obtained using the delta method (Oehlert 1992). In this simulation study, the coverage probability is defined as the observed percentage of times that the confidence interval includes the respective parameter. The bias and MSE in the estimation of a parameter η are given, respectively, by,

$$\widehat{\text{Bias}}(\widehat{\eta}) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\eta}^{(i)} - \eta\right) \quad \text{and} \quad \widehat{\text{MSE}}(\widehat{\eta}) = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\eta}^{(i)} - \eta\right)^2, \tag{26}$$

where we denote $\widehat{\eta}$ as each $\widehat{\alpha_1},\ \widehat{\alpha_2},\ \widehat{\beta}_1,\ \widehat{\beta}_2,\ \widehat{\rho}_1,\ \widehat{\rho}_2$ and $\widehat{\phi}$, η is the nominal value of the corresponded parameter, and N is the number of simulated samples of size n.

To generate bivariate data, we used an adaptation of the algorithm introduced by Balakrishnan and Lai (2009) and used by Ribeiro, Suzuki, and Saraiva (2017) and by Peres, Achcar, and Martinez (2018), along with an algorithm to defective distributions presented by Rocha, Nadarajah, Tomazella, Louzada, and Eudes (2017b) and used by Martinez and Achcar (2017, 2018). We generate random samples of size $n = 50, 75, 100, \ldots, 500$ in twelve different scenarios presented in Table (1). The steps of the proposed generation algorithm are described below.

Step 1: Fix values for the parameters: $\alpha_1,\ \alpha_2,\ \beta_1,\ \beta_2$ and ϕ.

Step 2: Calculate ρ_1 and ρ_2.

Step 3: Generate n random samples from $M_{1i} \sim Bernoulli(1-\rho_1)$.

Step 4: Generate n random samples from $u_{1i} \sim U(0, 1-\rho_1)$.

Step 5: For $i = 1, ..., n$ consider $t^*_{1i} = \infty$ if $M_{i1} = 0$ and $t^*_{1i} = F_1^{-1}(u_{1i})$ if $M_{i1} = 1$, where the inverse of the distribution function is given by,

$$F_1^{-1}(u_{1i}) = -\frac{1}{\beta_1}\ln\left[1 + \frac{\beta_1}{\alpha_1}\ln(1-u_{1i})\right]. \tag{27}$$

Step 6: Generate n random samples from $u^*_{1i} \sim U(0, max(t^*_{1i}))$, considering only finite values of t^*_{1i}.

Step 7: Consider $t_{1i} = min(t^*_{1i}, u^*_{1i})$.

Step 8. Pairs of values (t_{1i}, δ_{1i}) are thus obtained, where $\delta_{1i} = 1$ if $t_{1i} < u^*_{1i}$ and $\delta_{1i} = 0$ if $t_{1i} > u^*_{1i}$.

Step 9: Generate n random samples from $M_{2i} \sim Bernoulli(1-\rho_2)$.

Step 10: Generate n random samples from $u_{2i} \sim U(0, 1-\rho_2)$.

Step 11: Generate n random samples from $k_i \sim Bernoulli(\phi)$.

Step 12: Get values from w_i, considering the following expression,

$$w_i = \min\left\{u_{1i}^{-(\phi+1)}(u_{1i}^{\phi} + u_{2i}^{\phi})^{-\left(\frac{1+\phi}{\phi}\right)}, 1-\rho_2\right\}. \tag{28}$$

This expression is the derivative of (13) with respect to u_{i1}, when $S(t_{1i}) = u_{i1}$ and $S(t_{2i}) = w_i$.

Step 13: For $i = 1, ..., n$ consider $K_i = M_{1i}$ if $k_i = 1$ and $K_i = M_{2i}$ if $k_i = 0$.

Step 14: For $i = 1, ..., n$ consider $t^*_{2i} = \infty$ if $K_i = 0$ and $t^*_{2i} = F_2^{-1}(w_i)$ if $K_i = 1$, where the inverse of the distribution function is given by,

$$F_2^{-1}(u_{1i}) = -\frac{1}{\beta_2} \ln \left[1 + \frac{\beta_2}{\alpha_2} \ln(1 - w_i) \right]. \tag{29}$$

Step 15: Generate n random samples from $u^*_{2i} \sim U(0, max(t^*_{2i}))$, considering only finite values of t^*_{2i}.

Step 16: Consider $t_{2i} = min(t^*_{2i}, u^*_{2i})$.

Step 17. Pairs of values (t_{2i}, δ_{2i}) are thus obtained, where $\delta_{2i} = 1$ if $t_{2i} < u^*_{2i}$ and $\delta_{2i} = 0$ if $t_{2i} > u^*_{2i}$.

Table 1: Nominal values assumed for each scenario considered in the simulation study

		Scenarios											
		1	2	3	4	5	6	7	8	9	10	11	12
Parameter	ϕ	1.0				3.0				10.0			
	α_1	1.0	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0	0.5
	α_2	1.0	0.5	0.5	1.0	1.0	0.5	0.5	1.0	1.0	0.5	0.5	1.0
	β_1	0.8	1.5	0.8	1.5	0.8	1.5	0.8	1.5	0.8	1.5	0.8	1.5
	β_2	0.8	1.5	1.5	0.8	0.8	1.5	1.5	0.8	0.8	1.5	1.5	0.8

In the presence of a cure rate, it was considered nominal values for the parameters such that the samples generated have low and high percentage of cure rate in scenarios with parameter values $\alpha_i = 1.0$ and $\beta_i = 0.8$, where the cure rate parameter is given by $\rho_i \approx 0.2865$, and the scenarios parameter values $\alpha_i = 0.5$ and $\beta_i = 1.5$ where we have cure rate parameter given by $\rho_i \approx 0.7165$ $(i = 1, 2)$. From scenarios 1 to 4 (combinations of the fixed parameter values) in Table (1), it was considered $\phi = 1.0$, that is, $\tau_k(\phi) = 0.3333$ and $\tau_s(\phi) = 0.4790$, which corresponds to a moderate correlation between T_1 and T_2. from 5 to 8 (combinations of the fixed parameter values) given in Table (1), it was considered $\phi = 3.0$, so $\tau_k(\phi) = 0.6000$ and $\tau_k(\phi) = 0.7864$, representing a high correlation between T_1 and T_2. Finally, in the scenarios from 9 to 12 (combinations of the fixed parameter values), it was considered very high correlation between T_1 and T_2, with $\phi = 10.0$, which leads to $\tau_k(\phi) = 0.8333$ and $\tau_s(\phi) = 0.9583$ (see Table 1).

The ML estimates and corresponding standard errors for each simulated sample were computed using the *maxLik* package in *R* (Henningsen and Toomet 2011), and the Nelder-Mead maximization method, considering 95% nominal confidence intervals for the parameters. It was obtained in each scenario (Table (1)) the ML estimates of the parameters, the coverage probability of the confidence intervals, bias and MSE for each parameter of interest ρ_1, ρ_2 and ϕ, as well as the percentage of samples resulting in the presence of monotone likelihood functions (error informed by *maxLik*).

3.1. Results

This section presents simulations results, for each scenario presented in Table (1). It was observed that the percentage of censored data generated in the proposed simulation algorithm (Section 3) was about 5% higher than the respective percentage of the nominal cure rate (ρ_1 and ρ_2) considered in the generating samples. Moreover, for each simulated sample, it was calculated the Kendall's correlation τ_k by the Clayton copula approach and the Spearman correlation τ_s by numerical methods. Also, a re-parametrization of the parameter ϕ was considered in order to obtain flexible results for the coverage probability.

Figure (1) shows the box-plots of the ML estimates of the parameter ρ_1 in all scenarios considering different sample sizes (50 to 500), which enables us to observe the variability of these estimates. In each graph of Figure 1, horizontal dotted line refers to the nominal values of the parameter ρ_1. It is possible to see that the estimated values for ρ_1 are closer to the nominal vales, and the sampling variability decreases as the sample size increases as expected.

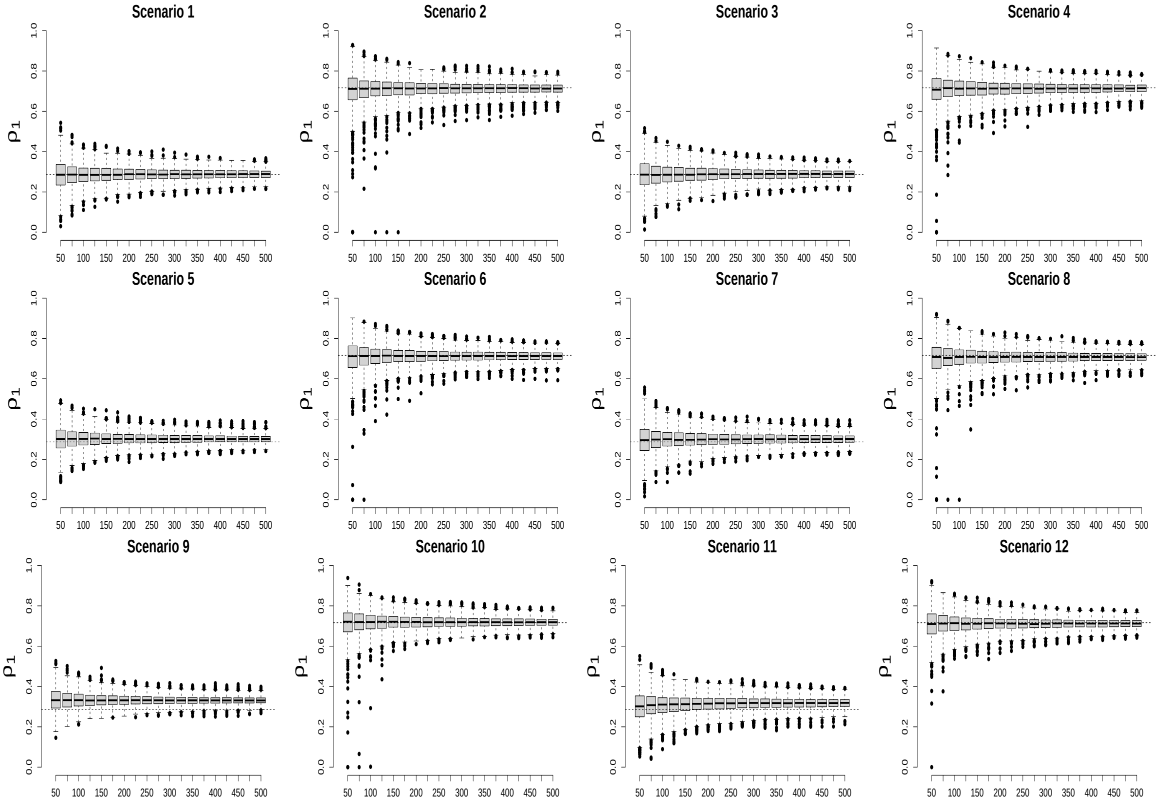

Figure 1: Box-plots of the maximum likelihood estimates for ρ_1 in each considered scenario considering different sample sizes

Figure (2) shows the box-plots of the ML estimates of the parameter ρ_2 in all considered scenarios, considering samples of size 50 to 500 in increments of 25. Comparing the results from Figures 1 and 2, we observe the presence of a higher bias for the estimates of ρ_2 than for the estimates of ρ_1, given that the estimated and the nominal values of ρ_1 and ρ_2 are not close to each other in scenarios 3, 4, 7, 8, 11 and 12. The higher biases and variability of the ML estimates for the parameter ρ_2 are observed in scenarios where we have high correlation between T_1 and T_2. Note that the estimates with lower biases are seen in the estimation of the parameter ρ_1 instead of the parameter ρ_2. This is probably due to the correlation between T_1 and T_2 included in the simulation process.

The box-plots for the estimates of ϕ are presented in the Figure (3). From these plots, it is possible to observe a great variability of the ML estimates of ϕ, and this variability increases as the correlation between T_1 and T_2 increases. Morever, it is also possible to observe relatively small interquartile ranges, indicating that most of the estimates obtained are highly concentrated in the central portion of the respective distributions, even in the presence of biases observed in the scenarios with higher cure rate. In addition, it is observed an expressive presence of bias in scenarios with higher cure rates, so that the medians of the estimates are slightly above the expected nominal values. In general, we could conclude that the model is adequate in these scenarios when the sample size is at least of 100 individuals.

Figures (4) and (5) shows the estimatives for the Kendall and Spearman correlation coefficients. Despite the difficulty to get an analytical expression for the Spearman correlation, it was obtained using numerical methods (see Section 2.2). Due to bias of the parameter ϕ (see Figure (3)) the Kendall and Spearman correlation measures were quite a bit higher than the

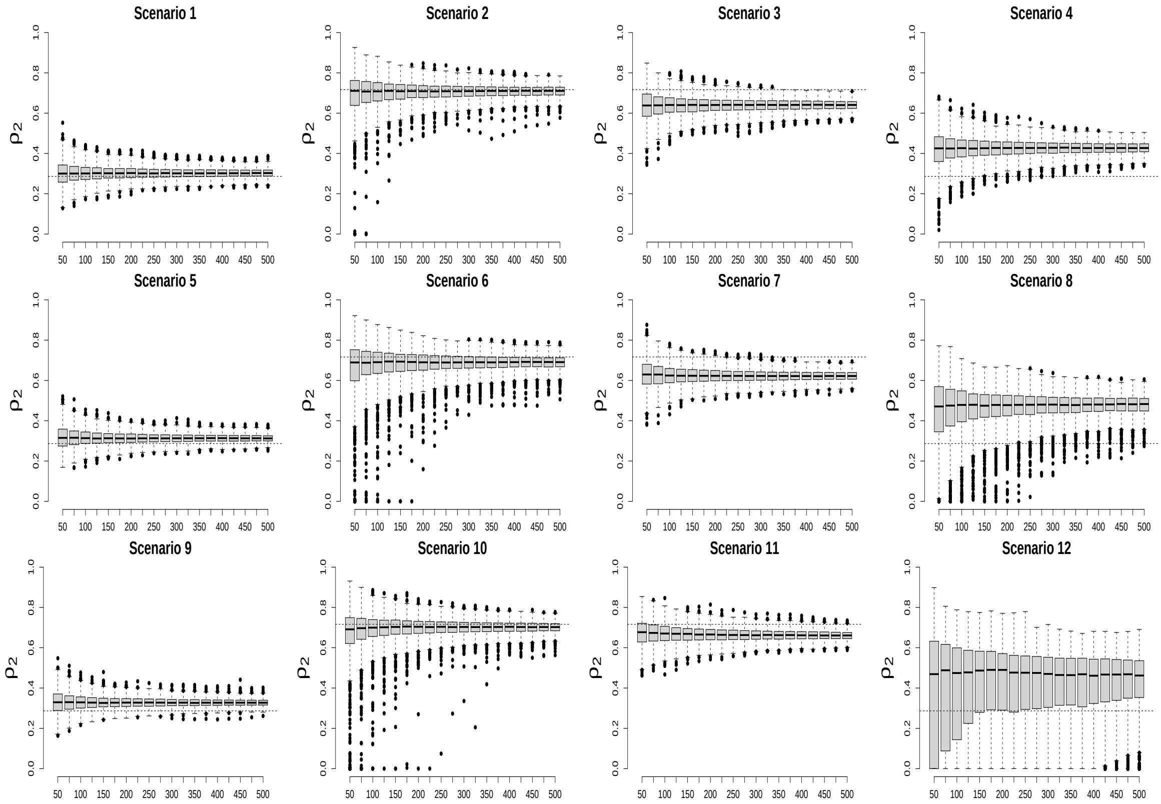

Figure 2: Box-plots of the maximum likelihood estimates for ρ_2 in each considered scenario considering different sample sizes

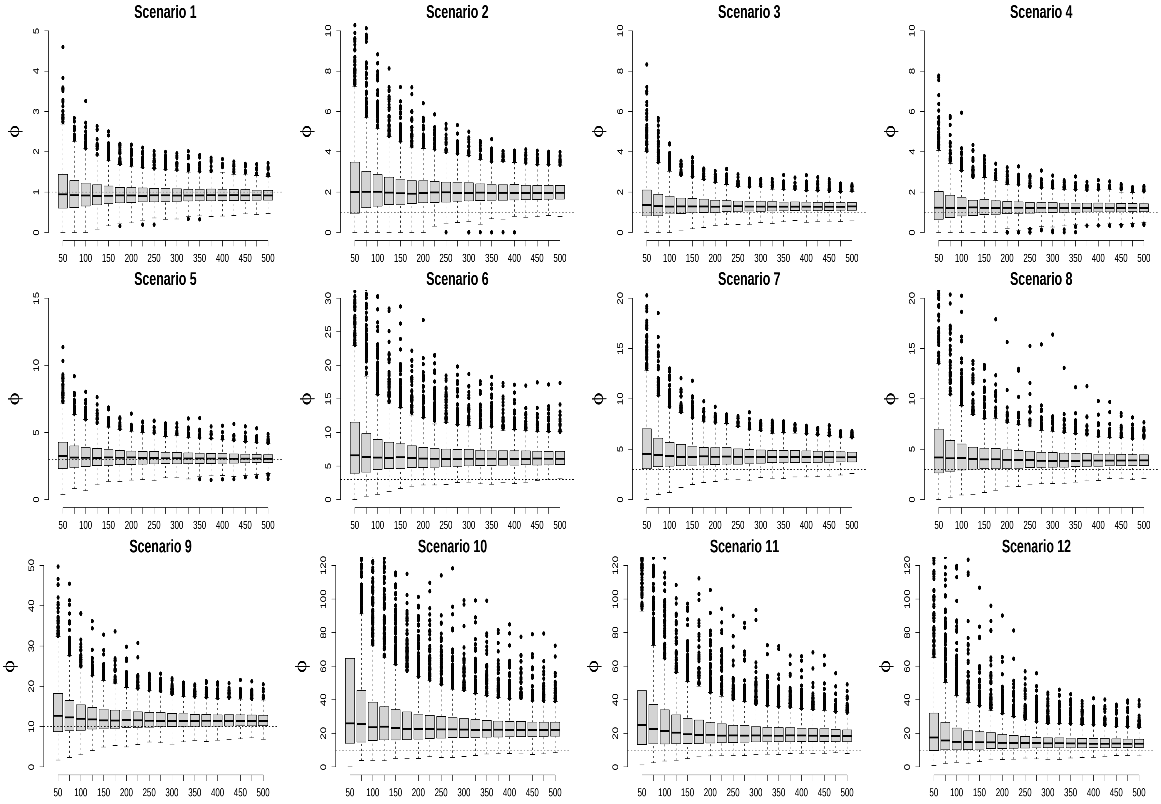

Figure 3: Box-plots of the estimates for ϕ, considering the maximum likelihood estimation method in different scenarios and different sample sizes

expected nominal values for these coefficients. These differences are identified mainly in the scenarios with higher cure rates and with high correlation between T_1 and T_2. In general, there is a great variability in the measurements obtained by Kendall and Spearman methods, despite the obtained results being close to the nominal values. However, this does not apply in situations where high Spearman correlation values between T_1 and T_2 are observed; in this case, almost all measurements are close to 1.

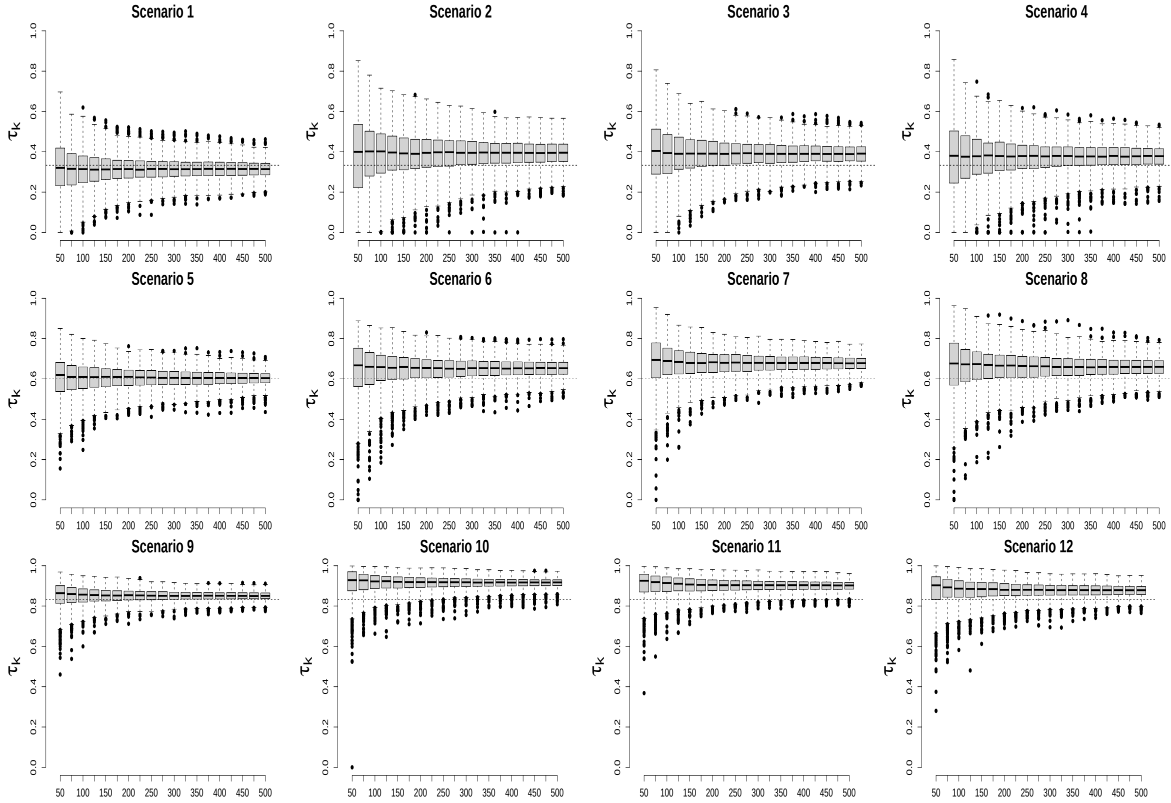

Figure 4: Box-plots of the estimates for τ_k, considering the maximum likelihood estimation method in different scenarios and different sample sizes

The confidence intervals may include or not the nominal values of the correspondent parameters. It was defined that the observed coverage probability is the number of times where the nominal value is inside to the corresponding confidence interval. This event can be modeled by a binomial distribution $Binomial(n, p)$, where n is number of simulated samples and p is the considered nominal coverage probability. In this paper, we used $n = 1000$ and $p = 0.95$ for each sample size used in the ML estimation, thus rejecting the equality between the nominal expected coverage probability and the observed coverage probability assuming a significance level of 5%, if the observed coverage probability is outside the range interval $(0.9365, 0.9635)$.

Figure (6) describes the coverage probability, bias and mean squared error for the parameter ϕ, in each considered scenario. Observing the graphs for $\phi = 1.0$ (low correlation between T_1 and T_2), the coverage probability is close to 95%. In the other scenarios the coverage probability in general it is greater than 95%. Besides that, it is possible to observe small biases, except in scenario 2, that considers a higher cure rate and produced a relatively high bias. The same does not apply to the cases $\phi = 3.0$ and $\phi = 10.0$. The scenarios 2 and 9 do not produce 95% coverage probability, and there are still large biases. As an important result, we can observe that when $\phi = 10.0$ the coverage probability is satisfactory for sample sizes larger than 300. Also it is observed that there are scenarios with high coverage probability when $\phi = 10.0$, however, this is due to the high estimated standard error for the parameter ϕ. In this case, there is the presence of high bias in the estimated value for the parameter ϕ, so the estimated range is not closed to the nominal value. This happens especially in scenarios with the presence of high cure rate in at least one of the time-to-event variables.

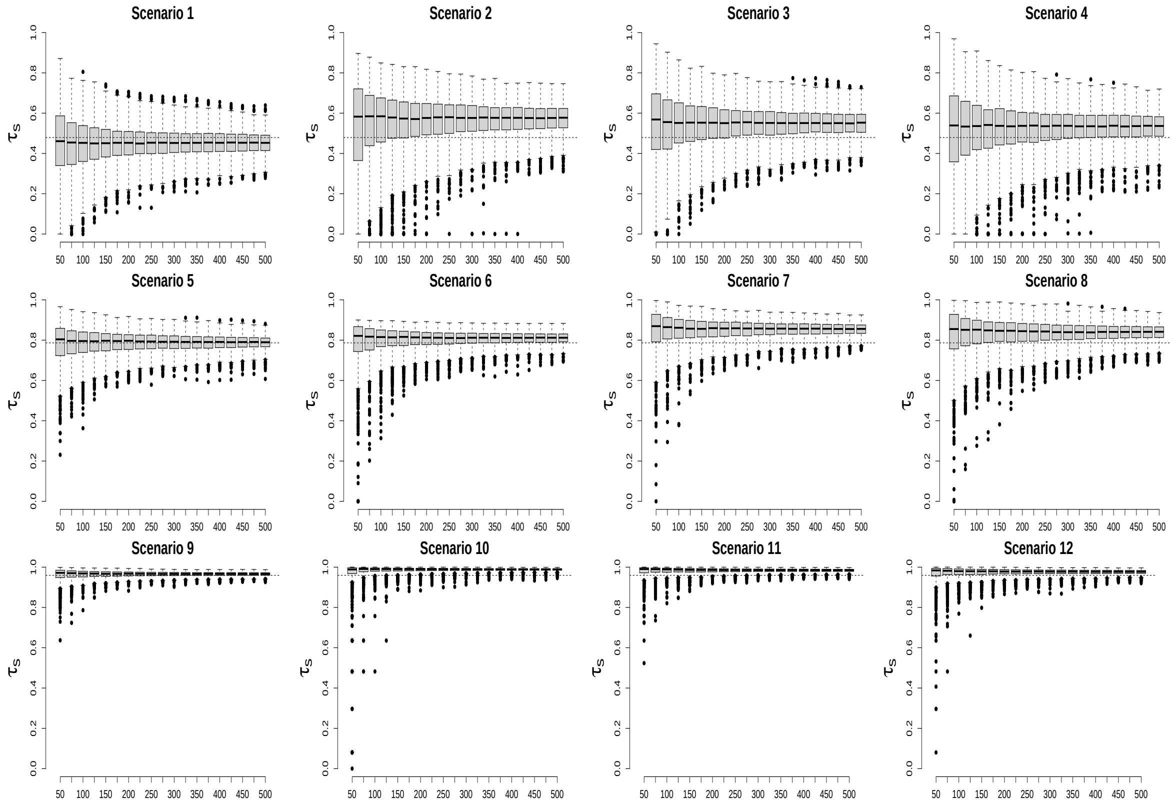

Figure 5: Box-plots of the estimates for τ_s, considering the maximum likelihood estimation method in different scenarios and different sample sizes

The standard errors for the estimates of the parameters ρ_1 and ρ_2 were calculated using the delta method, since these parameters are obtained as functions of the parameters α_1, α_2, β_1 and β_2. Figure (7) shows the coverage probability, bias and mean squared error for the cure rate parameter ρ_1, in all considered scenarios. The coverage probability has higher values in all assumed scenarios, given that the estimates for ρ_1 have lower bias and a relatively higher standard error. In these situations, probabilities are close to 100%. In addition, these results reinforce the conclusions previously obtained from Figure 1, where it is possible to conclude that the ML method adequately estimates the cure rate values.

Figure (8) shows the results of the simulation study considering the parameter ρ_2. The coverage probability is satisfactorily close to 95% in almost all scenarios. The same is not observed in the scenarios with different cure rates, in particular when the parameter ρ_1 is greater than parameter ρ_2, where we observed that the bias of the estimate for the parameter ρ_2 does not tend to 0, and the MSE is relatively high. This probably occurred due to the correlation considered in the simulation process of samples are pushing ρ_2 nearest to ρ_1. In all other scenarios, the bias of parameter ρ_2 is closest to 0.

It was noted during the simulation process, the presence of simulated samples that resulted in monotone likelihood functions, mainly when we considered samples sizes less than 200 and with high value for ϕ.

The coverage probability, biases and MSE for the estimators of the parameters α_1 α_2, β_1 and β_2 were also evaluated. Considering the scenarios with $\phi = 1.0$ and $\phi = 3.0$, the coverage probability, biases and MSE behave as expected, except for the estimator of the parameter β_2 that exhibit high bias and unexpected coverage probability. The parameters have different behaviors, reacting in different ways for each combination of parameter values, but it is noted that in the scenarios with low cure rate the bias are closer to zero. In addition, for all parameters, the bias and MSE decrease as the sample size increases, as it is expected.

It is observed in general a high bias related to the estimated parameters, and this bias is

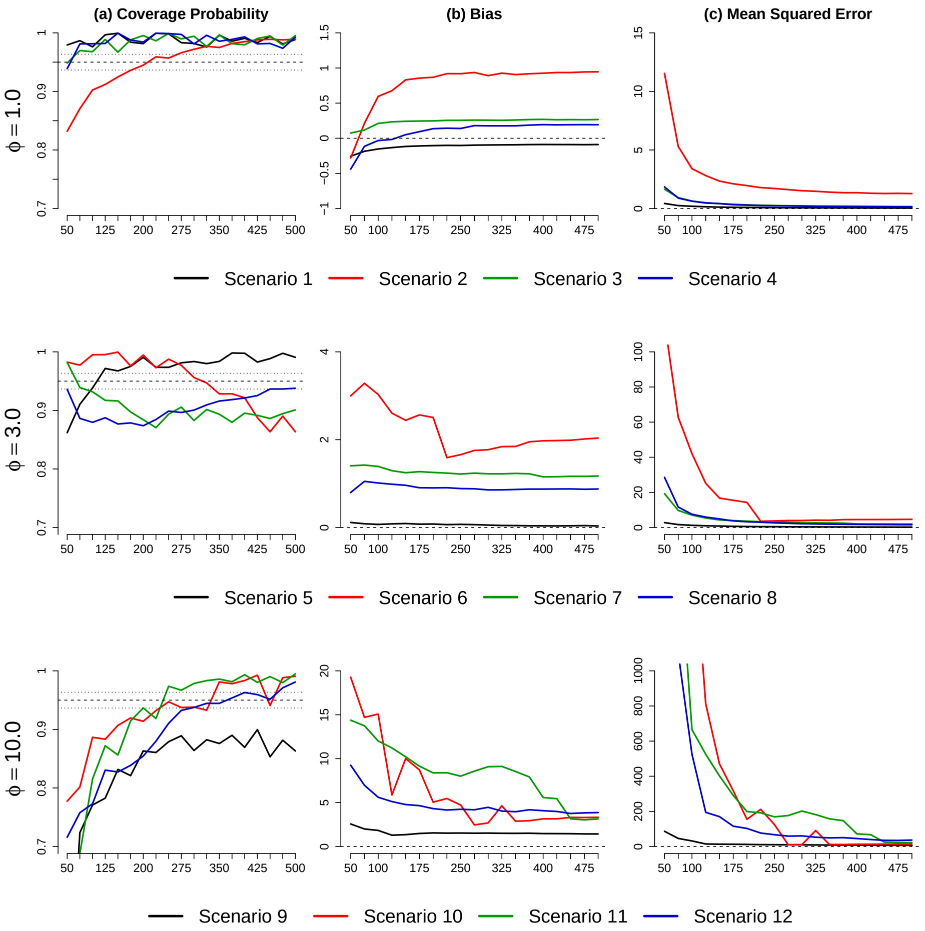

Figure 6: Plots of the coverage probability, biases and MSE for ϕ, considering the maximum likelihood estimation method

large enough to impair the probabilities of coverage of the correspondent confidence intervals. However, the range of bias were low compared to the parameter estimates. Probably, the previously mentioned problems associated with the parameter estimation are consequences of the method used to generate samples assuming a dependence structure between T_1 and T_2. In addition, the fit of BDGD bivariate model was verified for some samples by comparison of the estimated survival function with the Kaplan-Meier estimator. From these plots it was possible to see that the estimated survival curves by the BDGD model were satisfactorily closed to Kaplan-Meier curve, for both lifetimes T_1 and T_2. Besides, we can see in the graphs large dispersion, especially in small samples, which is probably due to problems of identifiability of the fraction curing models with mixtures, as described by Li, Taylor, and Sy (2001).

In a brief additional simulation study, it was considered a reparametrization for α_1, α_2, β_1 and β_2, where: $\gamma_k = \exp(\alpha_k)$ and $\lambda_k = \exp(\beta_k)$; also $\eta_k = \frac{1}{\alpha_k}$ and $\theta_k = \frac{1}{\beta_k}$, k=1,2. However, no significant changes were observed comparing the obtained inference results with the previously inference results presented in this section.

4. Applications to real data sets

In order to illustrate the proposed model, we present in this section four applications with real data sets. The first data set is related to breast cancer assuming 97 patients underwent surgical treatment for breast cancer followed up for a period between the year 2000 to 2011. The second dataset was introduced by Group *et al.* (1976) and is related to diabetic retinopathy disease. The third dataset is related to cervical cancer where it was observed two times: the disease-

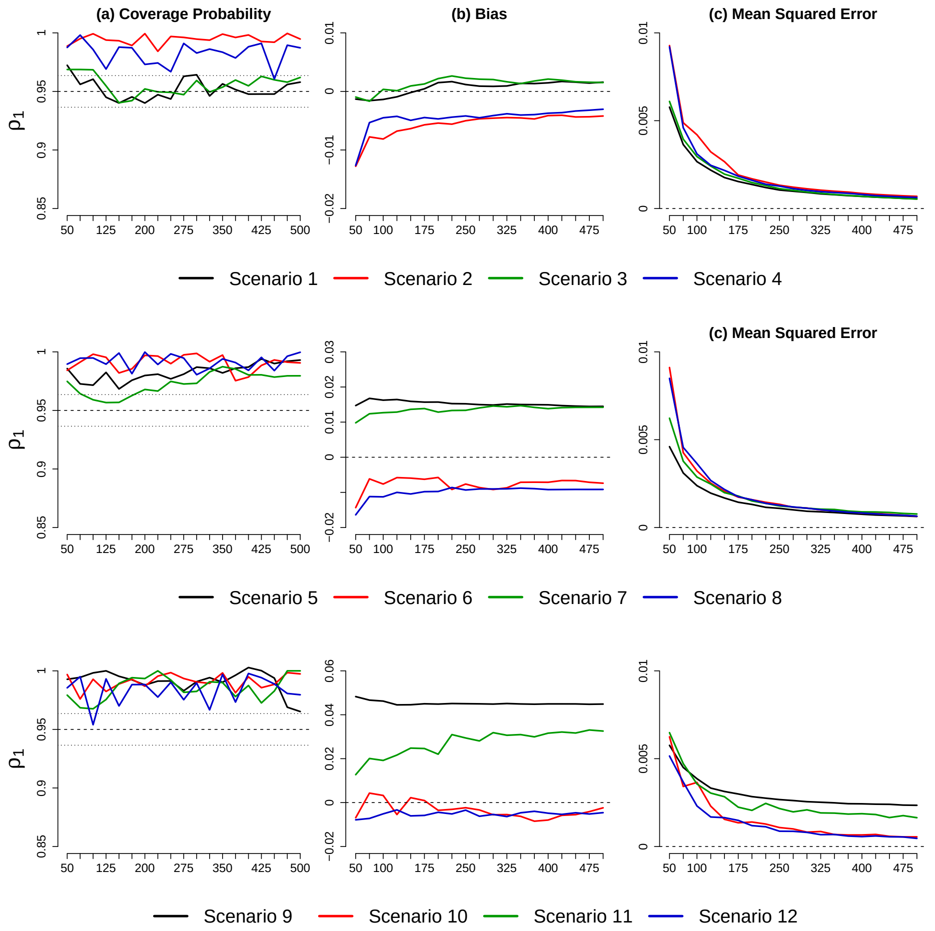

Figure 7: Plots of the coverage probability, biases and MSE for ρ_1, considering the maximum likelihood estimation method

free survival (DFS), defined as the time from the date of surgery to the first event of disease recurrence and the overall survival (OS), defined as the time from the date of surgery to the death. Finally, the fourth dataset is related to smokers patients that have tobacco-stained fingers. In each application, the Kendall correlation τ_k and the Spearman correlation τ_s, were compared with the empirical correlation between T_1 and T_2, denoted by τ_e, obtained by the package *SurvCorr* (Ploner, Kaider, and Heinze 2015). Also, it was compared the hazard function estimates by the proposed model with the empirical hazard function (obtained using the package “bshazard” Rebora, Salim, and Reilly 2018).

The obtained Bayesian estimates were based on 2,000 simulated Gibbs samples for the joint posterior distribution of interest recorded by every 50th iteration from 1,000,000 Gibbs samples after a “burn-in” period of 50,000 samples deleted to eliminate the effect of the initial values assumed in the simulation procedure. The convergence of the MCMC samples was checked by visual examination of traceplots of the simulated samples and convergence and stationary tests using the package *coda* Plummer, Best, Cowles, and Vines (2006). Approximately non-informative uniform prior distributions were assumed for the parameters of the BDGD model in almost all applications.

4.1. Application to a breast cancer data set

As first approach, it was considered a real dataset related to a cohort study, where 97 patients underwent surgical treatment for breast cancer followed up for a period between the year 2000 to 2011. For further details about the dataset, the reader should see Shigemizu, Iwase, Yoshimoto, Suzuki, Miya, Boroevich, Katagiri, Zembutsu, and Tsunoda (2017). For the

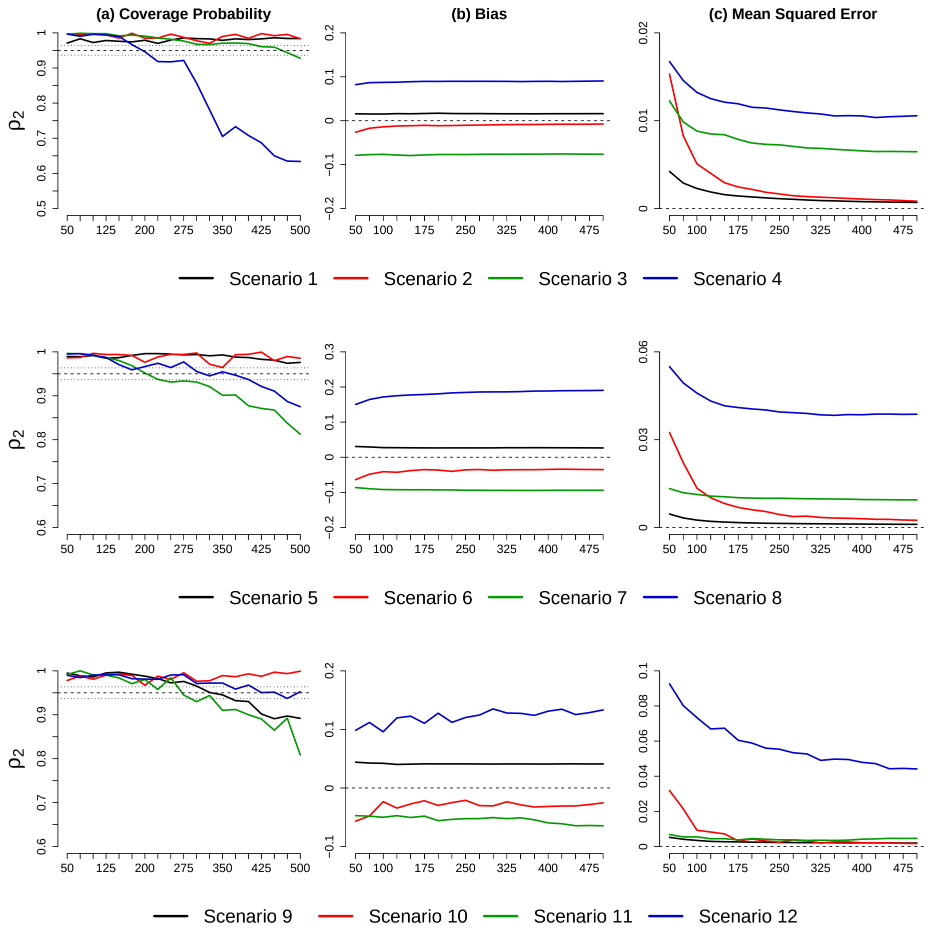

Figure 8: Plots of the coverage probability, biases and MSE for ρ_1, considering the maximum likelihood estimation method

bivariate lifetime application it was considered as T_1 the disease-free survival time (DFS) and T_2 representing the overall survival time (OS). In the dataset, there is 75% censored data for the disease-free survival time (T_1) and 80% censored data for the overall survival time (T_2).

Table (2) shows the estimates assuming the ML and the Bayesian methods for each parameter of the BDGD model assuming breast cancer data. Observe that the ML and Bayesian estimates are very close to each other, which means that the methods could be interchangeable. Moreover, the estimated values for τ_s obtained by copula functions are greater than the empirical correlation obtained by the R package *Survcorr* ($\tau_e = 0.8702\,(0.5288, 0.9691)$), but we can notice that the estimated value of τ_s is contained in the 95% confidence interval of τ_e, as shown in the simulation results.

The obtained Bayesian estimates for the parameters α_2 and β_2 are very close to the ML estimates, although there is a difference for the parameter ρ_2. This difference can be seen in Figure 9 showing the estimated survival (upper panels) and hazard (lower panels) curves for BDGD model proposed in this study. On the panel (b) of Figure 9, the survival curve estimated by a Bayesian approach diverges slightly from the Kaplan-Meier estimator for the survival function. In additional, on panel (d) of Figure 9, the hazard function estimated from ML approach is the closest to the empirical curve estimated by bshazard.

For T_1 both estimation methods resulted in similar curves. From the Kaplan-Meier plot for the survival function, it can be observed high cure rates in both lifetimes, where in T_1 there is a plateau close to the value 0.70, and in T_2 close to the value 0.75. These values are close to those estimated by the BDGD model. For the hazard curve, the model has a satisfactory fit to capture the decreasing shape of the empirical hazard function.

Table 2: Maximum likelihood estimates for the parameters of the BDGD model for the breast cancer data

Parameters	Maximum Likelihood Estimators			Bayesian Estimators	
	Estimate	Standard Error	95% CI	Median	95% CrI
α_1	0.1163	0.0305	(0.0858, 0.1470)	0.1139	(0.0326, 0.1486)
α_2	0.0576	0.0190	(0.0386, 0.0767)	0.0511	(0.0308, 0.0779)
β_1	0.2877	0.0853	(0.2025, 0.3731)	0.2747	(0.2032, 0.3910)
β_2	0.1980	0.0895	(0.1085, 0.2877)	0.2125	(0.1085, 0.2963)
ρ_1	0.6674	0.1096	(0.4525, 0.8823)	0.6752	(0.5143, 0.8943)
ρ_2	0.7474	0.1248	(0.5028, 0.9921)	0.7865	(0.5894, 0.8867)
ϕ	8.2022	2.0747	(4.1358, 12.2686)	7.8601	(4.1743, 11.7891)
τ_k	0.8039	0.0575	0.6912, 0.9167	0.7986	(0.6772, 0.8551)
τ_s	0.9431	-	-	0.9401	(0.8555, 0.9681)

95%CI: 95% confidence interval; 95%CrI: 95% credible interval.

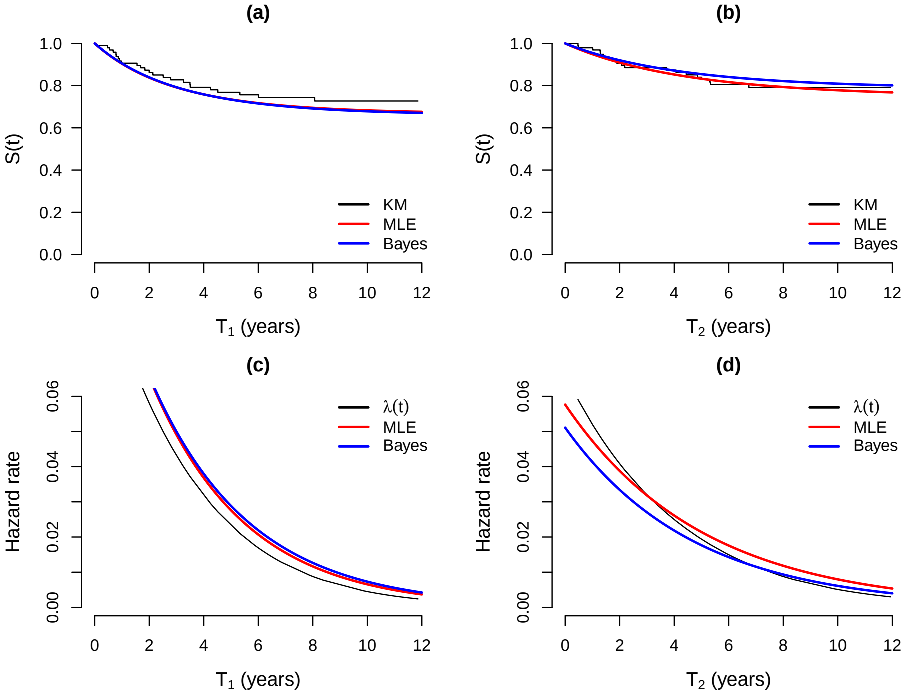

Figure 9: Plots of the survival functions estimated by Kaplan-Meier method and from the BDGD (upper panels) and respective hazard functions (lower panels) for DFS time (panels (a) and (c)) and OS time (panels (b) e (d)), considering breast cancer data

4.2. Application to a diabetic retinopathy data set

The diabetic retinopathy data used in the second application was introduced by Group *et al.* (1976). In this study, 197 diabetic patients patients up to 60 years old were followed-up for a fixed period. Each patient had one eye randomized for laser treatment and the other eye receiving no treatment. For a bivariate analysis T_1 is the time up to visual loss for the control eye, while T_2 corresponds to the time up to visual loss for the treatment eye. There was in this study 43% censored data of not treated eyes and 73% censored data of treated eyes.

Table 3 shows the estimates assuming the ML and the Bayesian methods for the parameters of the BDGD model assuming the diabetic retinopathy data. In this case, under both approaches, the estimate of the cure rate percentage are basically identical. Moreover, the Bayesian estimate for τ_s is smaller than the estimate obtained in the ML approach. However, the values of τ_s obtained from copula functions are significantly greater than the empirical correlation estimated by the *Survcorr* ($\tau_e = 0.3491\,(0.1071, 0.5522)$), and τ_s is contained in the 95% confidence interval for τ_e.

Figure 10 compares the survival curves $S(t_i)$ (upper panels) estimated from the Kaplan-Meier method and the empirical hazard function $\lambda(t_i)$ (lower panels), $i = 1, 2$, with the survival and hazard curves fitted by the BDGD model considering the retinopathy data. The ML estimates and Bayesian estimates produced similar plots. From the Kaplan-Meier estimator for the survival function, it was observed that the estimated fitted curves were very satisfactory for both T_1 and T_2. For the hazard curve, the model have a satisfactorily fit to capture the decreasing shape from the empirical hazard function. The estimated hazard curves using copula functions do not follow the total shape of the empirical hazard function for T_1 (panel (c)); it is possible that a more flexible distribution is needed for a better fitting. Also, it is important to say, that for this application it was needed to assume more informative prior distribution for the parameters $\beta_i (i = 1, 2)$ to get better convergence for the MCMC simulation algorithm.

Table 3: Maximum likelihood estimates for the parameters of the BDGD model for the retinopathy data

Parameters	Maximum Likelihood Estimators			Bayesian Estimators	
	Estimate	Standard Error	95% CI	Median	95% CrI
α_1	0.2781	0.0441	(0.2339, 0.3223)	0.2688	(0.2032, 0.3453)
α_2	0.1502	0.0325	(0.1178, 0.1828)	0.1486	(0.1022, 01973)
β_1	0.2239	0.0789	(0.1350, 0.2929)	0.2045	(0.1050, 0.2956)
β_2	0.3109	0.1104	(0.2005, 0.4214)	0.3277	(0.2068, 0.4442)
ρ_1	0.2725	0.1324	(0.0129, 0.5321)	0.2652	(0.0626, 04649)
ρ_2	0.6267	0.1262	(0.3693, 0.8642)	0.6342	(0.4372, 0.7708)
ϕ	0.9500	0.3479	(0.6021, 1.2979)	0.9038	(0.5228, 1.2805)
τ_k	0.3220	0.0379	(0.2476 0.3965)	0.3112	(0.2072, 0.3903)
τ_s	0.4634	-	-	0.4418	(0.3052, 0.5518)

95%CI: 95% confidence interval; 95%CrI: 95% credible interval.

4.3. Application to a cervical cancer data set

In this application, it is considered a medical data set from a published study by Brenna, Silva, Zeferino, Pereira, Martinez, and Syrjänen (2004) where it was also assumed the BDGD model. In this study 118 women received a standard treatment recommended to invasive cervical cancer. In a bivariate analysis T_1 is the disease-free survival (DFS), defined as the time from the date of surgery to the first event of disease recurrence and T_2 is the overall survival (OS), defined as the time from the date of surgery to the death. There is 48% censored data in T_1 and 53% censored data in T_2.

Table 4 presents the ML and Bayesian estimates for the parameters of the BDGD model assuming the cervical cancer data. In this case, the values of τ_s obtained by copulas functions are very close to the empirical correlation obtained by *Survcorr* ($\tau_e = 0.9118\,(0.8477, 0.9498)$), and the estimate ranges for τ_s estimated from Bayesian approach and τ_e are very similar. Peres *et al.* (2020) presented similar results for the correlation between T_1 and T_2 where the obtained value for the correlation between T_1 and T_2 was 0.8933.

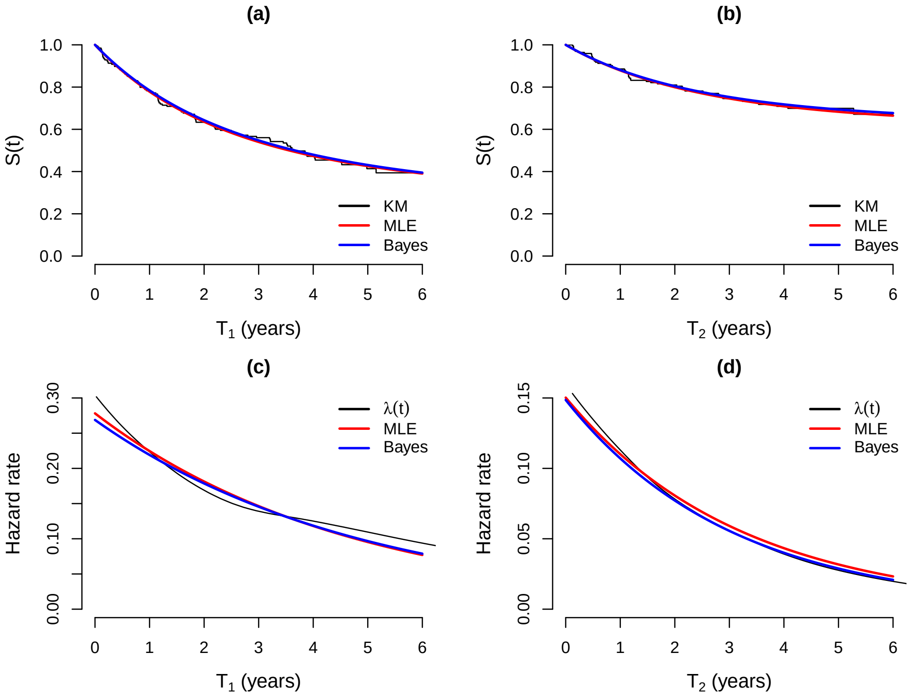

Figure 10: Plots of the survival functions estimated by Kaplan-Meier method and from the BDGD (upper panels) and respective hazard functions (lower panels) for control eye (panels (a) and (c)) and treatment eye (panels (b) and (d)), considering the diabetic retinopathy data

In general, the ML and Bayesian estimates produced almost identically plots for survival function and hazard function (Figure 11). Also, for the lifetime T_1, the fitted hazard based on the BDGD model has a slightly change from the empirical hazard curve obtained from package “bshazard” (panel (c)). However, the cure rate estimate for both times T_1 and T_2 was quite bad, that is, for T_2 the cure rate estimated by the ML and Bayesian approaches are 26% which is quite different from the 40% from the Kaplan-Meier curve. Finally, the hazard curves based on the BDGD fitting are close to the empirical hazard function (panel (d)).

Table 4: Maximum likelihood estimates for the parameters of the BDGD model for the cervical cancer data

Parameters	Maximum Likelihood Estimators			Bayesian Estimators	
	Estimate	Standard Error	95% CI	Median	95% CrI
α_1	0.4520	0.0775	(0.3744, 0.5296)	0.4431	(0.3547, 0.5541)
α_2	0.2060	0.0357	(0.1703, 0.2418)	0.1996	(0.1525, 0.2473)
β_1	0.4580	0.0859	(0.3721, 0.5440)	0.4524	(0.3559, 0.5451)
β_2	0.1537	0.0501	(0.1036, 0.2038)	0.1499	(0.1026, 0.1975)
ρ_1	0.3727	0.0935	(0.1894, 0.5561)	0.3736	(0.2463, 0.4961)
ρ_2	0.2617	0.1296	(0.0076, 0.5159)	0.2649	(0.1146, 0,4258)
ϕ	7.8998	1.2166	(5.5151, 10.2845)	8.0193	(5.1498, 10.8580)
τ_k	0.7979	0.0158	(0.7670 0.8290)	0.8003	(0.7202, 0.8444)
τ_s	0.9398	-	-	0.9411	(0.8890, 0.9635)

95%CI: 95% confidence interval; 95%CrI: 95% credible interval.

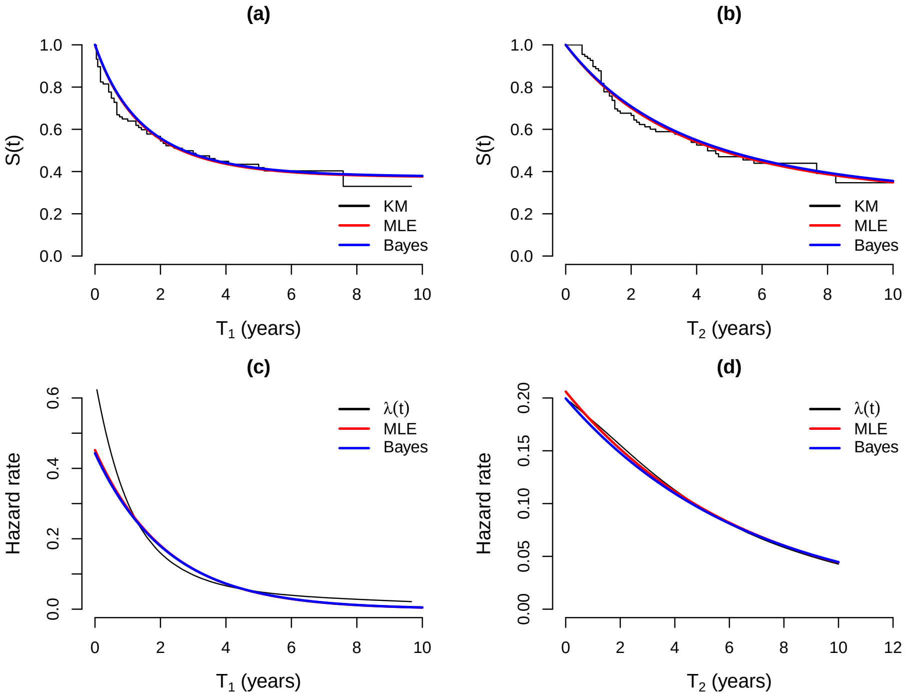

Figure 11: Plots of the survival functions estimated by Kaplan-Meier method and from the BDGD (upper panels) and respective hazard functions (lower panels) for DFS time (panels (a) and (c)) and OS time (panels (b) and (d)), considering the cervical cancer data

4.4. Application to a tobacco-stained fingers data set

As last application, we assumed a retrospective cohort study on a sample of 143 smokers screened between March 2006 and January 2010 in a 180-bed community hospital in La Chaux-de-Fonds, Switzerland. Data on death and hospital admission were collected until June 2014. More details on this data set can found in John, Louis, Berner, and Genné (2015). In this bivariate study, it is considered as T_1 the time before the first hospital readmission in smokers with stains on their fingers which was censored in case of death before the closure date;the lifetime T_2 is the survival time of the patient with tobacco-tar stain on their fingers. There was 26% censored data in T_1 and 48% censored data in T_2.

Table 5 shows the ML and Bayesian estimates for the parameters of the BDGD model considering the tobacco-stained fingers. In this application, it was needed assume more informative uniform prior distributions for the parameters α_1 and β_i $(i = 1, 2)$ for better inferences. In addition, the obtained estimates obtained considering the ML and Bayesian approach produced similar values, except for the parameter ϕ, where the Bayesian estimates were higher than the ML estimates. The value of τ_s obtained from ML estimates by copulas functions is equal to the empirical correlation ($\tau_e = 0.4998\,(0.2658, 0.6782)$). Bayesian estimates of τ_s are contained in the 95% confidence interval for τ_e. Similar results were obtained by de Oliveira *et al.* (2019).

Figure 12 shows the Kaplan-Meier plot from where we can notice that the T_1 has low cure rate and T_2 shows moderate cure rate (upper panels). The survival and hazard curves produced by ML estimates and Bayesian estimates are very similar. Based on the Kaplan-Meier curve for the empirical survival function, it is noted that the estimated curves were satisfactory fitted for T_1 and T_2. The BDGD model adequately estimated a cure rate in both ML and Bayesian approaches. Observing the estimated hazard function, we see that the proposed model captures in a good way , the decreasing shape of the hazard function. However, the hazard function estimated for T_2 based on the proposed BDGD model does not fully follow the behavior of the hazard function obtained from the package “bshazard” (panel (d)). It may be needed a more flexible distribution for a perfect fit of the hazard function in T_2.

Table 5: Maximum likelihood estimates for the parameters of the BDGD model for the tobacco data

Parameters	Maximum Likelihood Estimators			Bayesian Estimators	
	Estimate	Standard Error	95% CI	Median	95% CrI
α_1	0.6662	0.0923	(0.5739, 0.7585)	0.6648	(0.5095, 0.7937)
α_2	0.1990	0.0376	(0.1614, 0.2367)	0.2062	(0.1532, 02479)
β_1	0.3141	0.0715	(0.2426, 0.3856)	0.3242	(0.2538, 0.3962)
β_2	0.2444	0.0625	(0.1819, 0.3070)	0.2500	(0.1549, 0.3450)
ρ_1	0.1199	0.0590	(0.0041, 0.2357)	0.1303	(0.0552, 0.2445)
ρ_2	0.4428	0.1148	(0.2177, 0.6681)	0.4419	(0.2381, 0.6075)
ϕ	1.0477	0.2795	(0.4998, 1.5955)	1.3635	(0.5574, 7.9728)
τ_k	0.3437	0.0315	(0.2820 0.4056)	0.3889	(0.2179, 0.4965)
τ_s	0.4921	-	-	0.5463	(0.3204, 0.6784)

95%CI: 95% confidence interval; 95%CrI: 95% credible interval.

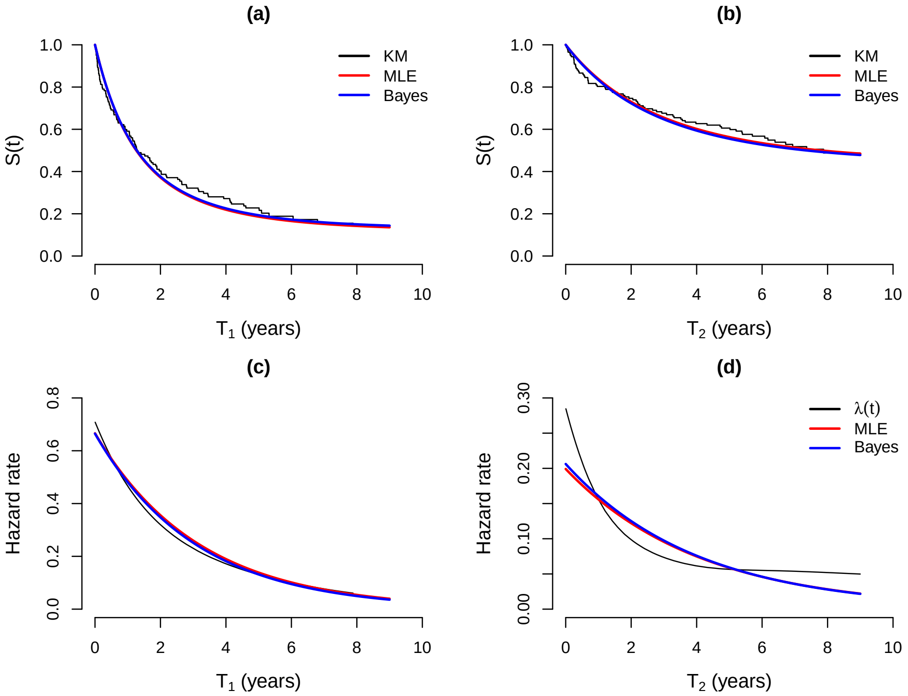

Figure 12: Plots of the survival functions estimated by Kaplan-Meier method and from the BDGD (upper panels) and respective hazard functions (lower panels) for first time hospital readmission (panels (a) and (c)) and survival time (panels (b) and (d)), considering the tobacco-stained fingers data

5. Concluding remarks

A new bivariate lifetime distribution model was proposed in this paper based on a defective Gompertz distribution and using the Clayton copula function in presence of cure fraction. Assuming the proposed new model, we performed a comprehensive simulation study to describe the performance of the inference results under the ML approach. This model is efficient to fit data with weak and strong correlation between lifetimes the T_1 and T_2 in several scenarios. However, in the situations where there is a high proportion of cure fraction and small sample sizes ($n < 100$)a careful use of this model is required. It was observed that the estimates are more easily obtained if the lifetime variables have values lower than 20, which demands some transformation in the data in some applications. In the application studies it was verified that both the ML and Bayesian methods are suitable approaches to estimate the parameters of the BDGD model. It is important to point out that a suitable choice for the initial values in the ML iterative estimation procedure is required, as well for the Bayesian method that depends on adequate hyperparameter values for the prior probability distributions for the parameters of the BDGD model. In conclusion, the applications in simulated and real data evidenced that the BDGD model can be satisfactorily fitted in most cases, considering both the ML and Bayesian approaches. Moreover, the proposed model can be easily implemented using *R* or Rjags softwares, which is a great advantage.

References

Achcar JA, Coelho-Barros EA, Mazucheli J (2012). "Cure Fraction Models Using Mixture and Non-mixture Models." *Tatra Mountains Mathematical Publications*, **51**(1), 1–9. doi:10.2478/v10127-012-0001-4.

Achcar JA, Coelho-Barros EA, Mazucheli J (2013). "Block and Basu Bivariate Lifetime Distribution in the Presence of Cure Fraction." *Journal of Applied Statistics*, **40**(9), 1864–1874. doi:10.1080/02664763.2013.798630.

Achcar JA, Martinez EZ, Tovar Cuevas JR (2016). "Bivariate Lifetime Modelling Using Copula Functions in Presence of Mixture and Non-mixture Cure Fraction Models, Censored Data and Covariates." *Model Assisted Statistics and Applications*, **11**(4), 261–276. doi:10.3233/MAS-160372.

Balakrishnan N, Lai CD (2009). *Continuous Bivariate Distributions*. Springer Vergland. ISBN 978-0387096148.

Balka J, Desmond AF, McNicholas PD (2011). "Bayesian and Likelihood Inference for Cure Rates Based on Defective Inverse Gaussian Regression Models." *Journal of Applied Statistics*, **38**(1), 127–144. doi:10.1080/02664760903301127.

Block HW, Basu AP (1974). "A Continuous, Bivariate Exponential Extension." *Journal of the American Statistical Association*, **69**(348), 1031–1037. doi:10.2307/2286184.

Boag JW (1949). "Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy." *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(1), 15–53. doi:10.2307/2983694.

Brenna SMF, Silva IDCG, Zeferino LC, Pereira JS, Martinez EZ, Syrjänen KJ (2004). "Prognostic Value of P53 Codon 72 Polymorphism in Invasive Cervical Cancer in Brazil." *Gynecologic Oncology*, **93**(2), 374–380. doi:10.1016/j.ygyno.2004.03.004.

Cancho VG, Bolfarine H (2001). "Modeling the Presence of Immunes by Using the Exponentiated-Weibull Model." *Journal of Applied Statistics*, **28**(6), 659–671. doi:10.1080/02664760120059200.

Cantor AB, Shuster JJ (1992). "Parametric versus Non-parametric Methods for Estimating Cure Rates Based on Censored Survival Data." *Statistics in Medicine*, **11**(7), 931–937. doi:10.1002/sim.4780110710.

Carlin BP, Louis TA (2000). "Empirical Bayes: Past, Present and Future." *Journal of the American Statistical Association*, **95**(452), 1286–1289. doi:10.2307/2669771.

Castro Md, Cancho VG, Rodrigues J (2009). "A Bayesian Long-term Survival Model Parametrized in the Cured Fraction." *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **51**(3), 443–455. doi:10.1002/bimj.200800199.

Chen MH, Ibrahim JG, Sinha D (1999). "A New Bayesian Model for Survival Data with a Surviving Fraction." *Journal of the American Statistical Association*, **94**(447), 909–919. doi:10.2307/2670006.

Chen MH, Ibrahim JG, Sinha D (2002). "Bayesian Inference for Multivariate Survival Data with a Cure Fraction." *Journal of Multivariate Analysis*, **80**(1), 101–126. doi:10.1006/jmva.2000.1975.

Chib S, Greenberg E (1995). "Understanding the Metropolis-Hastings Algorithm." *The American Statistician*, **49**(4), 327–335. doi:10.2307/2684568.

Clayton DG (1978). "A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence." *Biometrika*, **65**(1), 141–151. doi:10.1093/biomet/65.1.141.

Coelho-Barros EA, Achcar JA, Mazucheli J (2016). "Bivariate Weibull Distributions Derived from Copula Functions in the Presence of Cure Fraction and Censored Data." *Journal of Data Science*, **14**(2). URL http://www.jds-online.com/volume-14-number-2-april-2016.

Cook RD, Johnson ME (1981). "A Family of Distributions for Modelling Non-elliptically Symmetric Multivariate Data." *Journal of the Royal Statistical Society: Series B (Methodological)*, **43**(2), 210–218. doi:10.2307/2984851.

Corbière F, Commenges D, Taylor JM, Joly P (2009). "A Penalized Likelihood Approach for Mixture Cure Models." *Statistics in Medicine*, **28**(3), 510–524. doi:10.1002/sim.3481.

da Rocha RF, Tomazella LD, Louzada F (2014). "Bayesian and Classic Inference for the Defective Gompertz Cure Rate Model." *Revista Brasileira de Biometria*, **32**(1), 104–114.

De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A (1999). "Mixture Models for Cancer Survival Analysis: Application to Population-based Data with Covariates." *Statistics in Medicine*, **18**(4), 441–454. doi:10.1002/(sici)1097-0258(19990228)18:4<441::aid-sim23>3.0.co;2-m.

de Oliveira RP, Achcar JA, Peralta D, Mazucheli J (2019). "Discrete and Continuous Bivariate Lifetime Models in Presence of Cure Rate: A Comparative Study under Bayesian Approach." *Journal of Applied Statistics*, **46**(3), 449–467. doi:10.1080/02664763.2018.1495701.

dos Santos MR, Achcar JA, Martinez EZ (2017). "Bayesian and Maximum Likelihood Inference for the Defective Gompertz Cure Rate Model with Covariates: An Application to the Cervical Carcinoma Study." *Ciência e Natura*, **39**(2). doi:10.5902/2179460X24118.

Fachini JB, Ortega EMM, Cordeiro GM (2014). "A Bivariate Regression Model with Cure Fraction." *Journal of Statistical Computation and Simulation*, **84**(7), 1580–1595. doi:10.1080/00949655.2012.755531.

Gelfand AE, Smith AF (1990). "Sampling-based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, **85**(410), 398–409. doi:10.2307/2289776.

Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC Press. ISBN 978-0203491287.

Ghitany ME, Maller RA (1992). "Asymptotic Results for Exponential Mixture Models with Long-term Survivors." *Statistics: A Journal of Theoretical and Applied Statistics*, **23**(4), 321–336. doi:10.1080/02331889208802379.

Gieser PW, Chang MN, Rao PV, Shuster JJ, Pullen J (1998). "Modelling Cure Rates Using the Gompertz Model with Covariate Information." *Statistics in Medicine*, **17**(8), 831–839. doi:10.1002/(sici)1097-0258(19980430)17:8<831::aid-sim790>3.0.co;2-g.

Group DRSR, *et al.* (1976). "Preliminary Report on Effects of Photocoagulation Therapy." *American Journal of Ophthalmology*, **81**(4), 383–396. doi:10.1016/j.ajo.2017.11.010.

Henningsen A, Toomet O (2011). "maxLik: A Package for Maximum Likelihood Estimation in R." *Computational Statistics*, **26**(3), 443–458. doi:10.1007/s00180-010-0217-1.

Hofert M, Kojadinovic I, Mächler M, Yan J (2019). *Elements of Copula Modeling with R*. Springer. ISBN 978-3319896342.

Joe H (2014). *Dependence Modeling with Copulas*. Chapman and Hall/CRC Press. ISBN 978-1466583221.

John G, Louis C, Berner A, Genné D (2015). "Tobacco Stained Fingers and Its Association with Death and Hospital Admission: A Retrospective Cohort Study." *PLOS ONE*, **10**(9), e0138211. doi:10.1371/journal.pone.0138211.

Kleinbaum DG, Klein M (2012). *Survival Analysis: A Self-Learning Text*. Springer. ISBN 978-0387239187.

Lambert PC, Thompson JR, Weston CL, Dickman PW (2006). "Estimating and Modeling the Cure Fraction in Population-based Cancer Survival Analysis." *Biostatistics*, **8**(3), 576–594. doi:10.1093/biostatistics/kxl030.

Li CS, Taylor JMG, Sy JP (2001). "Identifiability of Cure Models." *Statistics & Probability Letters*, **54**(4), 389–395. doi:10.1016/s0167-7152(01)00105-5.

Li Y, Tiwari RC, Guha S (2007). "Mixture Cure Survival Models with Dependent Censoring." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(3), 285–306. doi:10.2307/4623270.

Maller RA, Zhou X (1996). *Survival Analysis with Long-term Survivors*. Wiley New York. ISBN 978-0471962014.

Marshall AW, Olkin I (1967). "A Generalized Bivariate Exponential Distribution." *Journal of Applied Probability*, **4**(2), 291–302. doi:10.1017/s0021900200032058.

Martinez EZ, Achcar JA (2014). "Bayesian Bivariate Generalized Lindley Model for Survival Data with a Cure Fraction." *Computer Methods and Programs in Biomedicine*, **117**(2), 145–157. doi:10.1016/j.cmpb.2014.07.011.

Martinez EZ, Achcar JA (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**(2), 463–484. doi:10.1285/i20705948v10n2p463.

Martinez EZ, Achcar JA (2018). "A New Straightforward Defective Distribution for Survival Analysis in the Presence of a Cure Fraction." *Journal of Statistical Theory and Practice*, **12**(4), 688–703. doi:10.1080/15598608.2018.1460885.

Martinez EZ, Achcar JA, Jácome AAA, Santos JS (2013). "Mixture and Non-mixture Cure Fraction Models Based on the Generalized Modified Weibull Distribution with an Application to Gastric Cancer Data." *Computer Methods and Programs in Biomedicine*, **112**(3), 343–355. doi:10.1016/j.cmpb.2013.07.021.

Nelsen RB (2007). *An Introduction to Copulas*. Springer Science & Business Media. ISBN 978-0387286594.

Oakes D (1982). "A Model for Association in Bivariate Survival Data." *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(3), 414–422. doi:10.2307/2345500.

Oehlert GW (1992). "A Note on the Delta Method." *The American Statistician*, **46**(1), 27–29. doi:10.2307/2684406.

Peres MVdO, Achcar JA, Martinez EZ (2018). "Bivariate Modified Weibull Distribution Derived from Farlie-Gumbel-Morgenstern Copula: A Simulation Study." *Electronic Journal of Applied Statistical Analysis*, **11**(2), 463–488. doi:10.1285/i20705948v11n2p463.

Peres MVdO, Achcar JA, Martinez EZ (2020). "Bivariate Lifetime Models in Presence of Cure Fraction: A Comparative Study with Many Different Copula Functions." *Heliyon*, **6**(6), e03961. doi:10.1016/j.heliyon.2020.e03961.

Ploner M, Kaider A, Heinze G (2015). *SurvCorr: Correlation of Bivariate Survival Times*. R package version 1.0, URL https://CRAN.R-project.org/package=SurvCorr.

Plummer M, Best N, Cowles K, Vines K (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. URL https://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf.

Plummer M, *et al.* (2003). "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In *Proceedings of the 3rd international workshop on distributed statistical computing*, pp. 1–10. Vienna, Austria. URL https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf.

Rebora P, Salim A, Reilly M (2018). *bshazard: Nonparametric Smoothing of the Hazard Function*. R package version 1.1, URL https://CRAN.R-project.org/package=bshazard.

Ribeiro TR, Suzuki AK, Saraiva EF (2017). "Uma Abordagem Bayesiana para o Modelo de Sobrevivência Bivariado Derivado da Cópula AMH." *Revista da Estatística da Universidade Federal de Ouro Preto*, **6**(1), 1–20. URL https://periodicos.ufop.br:8082/pp/index.php/rest/article/view/3343.

Rocha R, Nadarajah S, Tomazella V, Louzada F (2017a). "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. doi:10.1016/j.csda.2016.10.001.

Rocha R, Nadarajah S, Tomazella V, Louzada F, Eudes A (2017b). "New Defective Models Based on the Kumaraswamy Family of Distributions with Application to Cancer Data Sets." *Statistical Methods in Medical Research*, **26**(4), 1737–1755. doi:10.1177/0962280215587976.

Schemper M, Kaider A, Wakounig S, Heinze G (2013). "Estimating the Correlation of Bivariate Failure Times under Censoring." *Statistics in Medicine*, **32**(27), 4781–4790. doi:10.1002/sim.5874.

Shigemizu D, Iwase T, Yoshimoto M, Suzuki Y, Miya F, Boroevich KA, Katagiri T, Zembutsu H, Tsunoda T (2017). "The Prediction Models for Postoperative Overall Survival and Disease-free Survival in Patients with Breast Cancer." *Cancer Medicine*, **6**(7), 1627–1638. doi:10.1002/cam4.1092.

Tsodikov AD, Ibrahim JG, Yakovlev AY (2003). "Estimating Cure Rates from Survival Data: An Alternative to Two-component Mixture Models." *Journal of the American Statistical Association*, **98**(464), 1063–1078. doi:10.2307/30045351.

Vahidpour M (2016). *Cure Rate Models*. Ph.D. thesis, École Polytechnique de Montréal. URL https://publications.polymtl.ca/2454/.

Vaupel JW, Manton KG, Stallard E (1979). "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography*, **16**(3), 439–454. doi:10.2307/2061224.

Wienke A (2010). *Frailty Models in Survival Analysis*. CRC press. ISBN 978-1420073911.

Wienke A, Lichtenstein P, Yashin AI (2003). "A Bivariate Frailty Model with a Cure Fraction for Modeling Familial Correlations in Diseases." *Biometrics*, **59**(4), 1178–1183. doi:10.2307/3695360.

Wienke A, Locatelli I, Yashin AI (2006). "The Modelling of a Cure Fraction in Bivariate Time-to-event Data." *Austrian Journal of Statistics*, **35**(1), 67–76. doi:10.17713/ajs.v35i1.349.

Yu B, Peng Y (2008). "Mixture Cure Models for Multivariate Survival Data." *Computational Statistics & Data Analysis*, **52**(3), 1524–1532. doi:10.1016/j.csda.2007.04.018.

Affiliation:

Marcos Vinicius de Oliveira Peres
Ribeirão Preto Medical School
University of São Paulo (USP)
Ribeirão Preto, SP, Brazil
E-mail: mvperes1991@alumni.usp.br

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 51 *Submitted:* 2020-12-30
January 2022 *Accepted:* 2021-03-30