Instituto de
MATEMÁTICA
E ESTATÍSTICA
UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Modelo de fração de cura para censura dependente sob abordagem de cópulas

Maicon Gottselig

3 de junho de 2021

Maicon Gottselig

Modelo de fração de cura para censura dependente sob abordagem de cópulas

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Profa. Dr. Silvana Schneider

Porto Alegre
Maio de 2021

Maicon Gottselig

Modelo de fração de cura para censura dependente sob abordagem de cópulas

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a):________________________________
Profa. Dr. Silvana Schneider, UFMG
Doutor(a) pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Banca Examinadora:

Prof. Dr. Guilherme Pumi, UFRGS
Doutor(a) pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Porto Alegre
Maio de 2021

Agradecimentos

Agradeço aos meus familiares por verem em mim coisas que às vezes esqueço, por acreditarem em momentos que questiono e por confortarem quando anseio. Obrigado por me permitirem chegar aqui.

Agradeço principalmente aos meus avós. Lori, mulher forte, corajosa e altruísta, me concedeu as asas com que voo e as raízes para quais volto. Paulo, com suas palavras muito aprendi e através de você muito realizei.

Sou grato aos meus amigos pela colaboração nesta trajetória, agradeço pelos ombros, ouvidos, conselhos e risadas.

Este momento só é possível por causa dos professores do departamento de estatística, obrigado pela condução. Obrigado especial à minha orientadora Silvana, esta etapa final seria impossível sem sua confiança, paciência e motivação.

Por último, agradeço à Universidade Federal do Rio Grande do Sul pelo acolhimento e transformação que pude viver.

Resumo

Este trabalho é centrado na formulação de uma abordagem de cura para censura dependente sob abordagem de funções cópulas. O modelo de não-mistura é considerado para permitir fração de cura. A dependência entre o tempo até evento de interesse e tempo de censura dependente é ajustada por meio de funções cópula. São apresentadas a função de verossimilhança do modelo proposto e estimação de máxima verossimilhança. São assumidas as distribuições Weibull e exponencial por partes para ajuste dos tempos até evento de interesse e tempos de censura dependente. As funções cópulas de Clayton e de Plackett são utilizadas para ajuste da dependência. Um estudo de simulação foi conduzido para avaliar os modelos propostos, diferentes cenários de dependência foram assumidos para avaliar os efeitos nas estimativas do modelo. Um conjunto de dados sobre tempo de sobrevivência de pacientes diagnosticados com câncer de próstata é analisado com os modelos propostos.

Palavras-Chave: Análise de sobrevivência, Modelo de fração de cura, Tempo de promoção, Função cópula, Censura dependente, Câncer de próstata.

Abstract

This paper is centered around the formulation of a cure rate model for dependent censoring under copula functions. Non-mixture model is assumed to allow for a cure rate. Dependency between time to event if interest and dependent censoring time is accommodated via copula functions. The model's likelihood function, as well as, the maximum likelihood estimation, are presented. Weibull and piecewise exponential distribution are assumed to model time to event of interest and dependent censoring time. Clayton and Plackett copulas functions are used to capture dependency. A simulation study was conducted to evaluate the proposed models and different dependency scenarios were assumed to assess effects on the model's estimates. A real world dataset about prostate cancer patients survival time is analyzed with the proposed models proposed.

Keywords: Survival Analysis, Cure rate model, Promotion time, Copula function, Dependent censoring, Prostate cancer.

ARTIGO CIENTÍFICO

Modelo de fração de cura para censura dependente sob abordagem de cópulas

Maicon Gottselig, bacharelado em Estatística pela UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL(UFRGS)

CURE RATE MODEL FOR DEPENDENT CENSORING UNDER THE COPULA APPROACH

Maicon Gottselig[1]
Silvana Schneider[2]

- ABSTRACT: This paper is centered around the formulation of a cure rate model for dependent censoring under copula functions. Non-mixture model is assumed to allow for a cure rate. Dependency between time to event if interest and dependent censoring time is accommodated via copula functions. The model's likelihood function, as well as, the maximum likelihood estimation, are presented. Weibull and piecewise exponential distribution are assumed to model time to event of interest and dependent censoring time. Clayton and Plackett copulas functions are used to capture dependency. A simulation study was conducted to evaluate the proposed models and different dependency scenarios were assumed to assess effects on the model's estimates. A real world dataset about prostate cancer patients survival time is analyzed with the proposed models proposed.

- KEYWORDS: Survival Analysis; Cure rate model; Promotion time; Copula function; Dependent censoring; Prostate cancer.

1 Introduction

This paper aims to present the building of a cure rate model for dependent censoring under a copula-based approach, a solution to model survival data with a cure rate in the presence of dependent censoring. A cure rate model is necessary when the study population is composed of two groups, *susceptible* and *cured* individuals, when some observations are censored. *Susceptible* individuals can experience the event of interest if no censoring occurs. On the other hand, *cured* individuals will not experience the event, but may be censored (Klein et al., 2016). A censored observation happens when the event of interest cannot be observed due the

[1] Federal University of Rio Grande do Sul,Porto Alegre,RS,Brazil. E-mail: *maiconmfg@gmail.com*
[2] Federal University of Rio Grande do Sul,Porto Alegre,RS,Brazil. E-mail: *schneider.sil@gmail.com*

occurrence of some secondary event, such as end of study, patient withdrawal, failure due other causes (Lawless, 2011). Although most methods assume independence between the event of interest time and the censoring time, this assumption is true in some censoring cases, but not feasible in others.

Many cure rate models assume independence between time to event of interest and the censoring time (Wang et al., 2021). This offers simplification in the likelihood formulation because the joint distribution is not required (Kalbfleisch and Prentice, 2011). The incorrectly assumed independence leads to biased estimates and misleading conclusion over parameters (Siannis, 2004). There are many cases in which the independence assumption can be violated. According to William and Lagakos (1977), in clinical contexts, individuals withdrawing from a study due to reasons linked to the therapy under study can be a dependent censoring of time to event of interest. Alternatively, in competing risks, some risks might be correlated, as stated in Hsu et al. (2016).

Dependent censoring has been addressed in research. Approaches adjusting dependency include the frailty model in Huang and Wolfe (2002), that adjusted dependency using random effects, and Emura and Michimae (2017) that used copula functions, but they do not account for a cure rate. The omission of a cure rate can also lead to misleading inferences (Rondeau et al., 2013). Some cure rate models that account for dependent censoring are the following: Li et al. (2007), which used copulas functions to model dependent censoring in survival data without covariates; Zhang et al. (2007), which presented a frailty cure model where the covariates effects on the cure rate and on the event time of *susceptible* individuals are separately modeled; Liu et al. (2017), which formulated a regression model that accounts for dependent censoring and cure rate using latent variables; and Wang et al. (2021), which allowed a cure rate through a logistic model, using an additive hazards model with frailty, when interval dependent censoring were present.

In this paper we use copula functions to adjust for dependent censoring. Zheng and Klein (1995) developed the survival copula model, where the copula parameter adjusts the dependency between time to event of interest and censoring time. Zheng and Klein (1995) also explains that copula models are non-identifiable under certain specifications. Because of this, many authors present a sensitivity analysis (Chen, 2010; Emura and Michimae, 2017). Sensitivity analyses evaluate the effect of the dependency feature on parameters estimates (Huang and Zhang, 2008). Other works, such as Escarela and Carriere (2003), Hsu et al. (2016) and Li et al. (2019), estimate the copula parameter and prove their model's identifiability.

Copula functions are good options to adjust for dependent censoring because clustering is not required as in the frailty model. Copula functions are flexible and there are many copulas with various characteristics. Copula functions can also be extended to competing risk. This paper will study two particular copulas: the archimedean Clayton copula from Clayton (1978), and the non-archimedean Plackett copula from Plackett (1965).

The Clayton copula is a traditional copula choice. Its dependency structure is asymmetric, with a strong dependence in the lower tail and a weak dependence in the upper tail (Salvadori et al., 2007). Being an archimedean copula, it presents a generator function that allows easy extrapolation to higher dimensions (Nelsen, 2007).

The Plackett copula's parameter is related to the odds ratio association measure of 2 × 2 contingency tables. It is positive ordered and presents radial symmetry (Palaro and Hotta, 2006). The Plackett copula is not archimedean and so does not present a generator function (Salvadori et al., 2007).

Both copulas present parameters that relate to known dependence measures. The Clayton copula relates directly to Kendall's correlation coefficient τ, while Plackett copula is directly related to Spearman's correlation coefficient ρ. Analyzing the copula parameter's range and its relation to the known dependence measure we are able to conclude the correlation range the copula is able to adjust. For instance, Farlie-Gumbel-Morgenstern copula cannot adjust correlations with Kendall's $|\tau| > 0.33$ (Salvadori et al., 2007), or Galambos copula that can only adjust correlations with Kendall's $\tau > 0$ (Salvadori et al., 2007). Clayton and Plackett however do not present such correlation ranges restrictions. The Clayton copula needs some modifications to express negative dependence as showed in Emura and Chen (2018).

The marginal distributions of the copula model can be parametrically or non-parametrically specified, and these marginal distributions can depend on covariates and account for a cure rate. This paper will be present the formulation to model survival time, conditioned on covariates, and consider a cure rate that can also depend on covariates.

In order to model a cure rate, two main approaches exist. The first, proposed in Boag (1949), is held as the standard cure rate model and known as mixture model. The second class of cure rate model, which is the scope of this paper, is known as non-mixture cure rate model or promotion time cure rate model, derived of Yakovlev et al. (1993). It was formulated upon the number of active cancer cells. In both models, latent variables specify if each individual is *cured* or not. We will focus on the coditional approach to the non-mixture cure rate model.

The mixture cure rate model is a special case of the promotion time model. Medical community agrees that the promotion time model has more meaningful biological interpretation (Yakovlev et al., 1994). Some recent works include the following: Cancho et al. (2011), which used Bayesian approach to models cure rate using promotion time models with negative binomial distribution; Lambert and Bremhorst (2019), which studied the effect on using same covariates to model survival time and cure rate; Lambert and Bremhorst (2020), which included time-varying covariates to model cure rate; and Han et al. (2021), which proposed a semiparametric estimation method for the promotion time using auxiliary covariates to model cure rate.

In order to parametrically specify the marginal time to event of interest distribution we use Weibull and piecewise exponential distributions. Both distributions are common choices in survival analysis and continue to be researched, as seen in Almetwally et al. (2018). The study derives the maximum likelihood estimation and the Bayesian estimation to the Weibull generalized exponential distribution. On the other hand, Wey et al. (2020) use piecewise exponential distribution with time-varying effects to estimate mortality in organ transplant.

To assess the model's behavior we conduct a simulation study. We generated datasets from one model and adjusted them under different dependency scenarios. In order to see the model's applicability on real data we fit prostate cancer dataset from SEER (National Institutes of Health surveillance epidemiology and end results). Studies indicate that prostate cancer and cardiovascular disease survival time might be correlated (Escarela and Carriere, 2003; Li et al., 2007; Rowley et al., 2017; Cardwell et al., 2020). Additionally, due to recent medical developments prostate cancer patients present high cure chances.

This paper is organized as follows. Section 2 presents definitions of the non-mixture method to allow cure rate, the formulation of the marginal distributions of time to event of interest and censoring time, the construction of the likelihood functions, and Weibull and piecewise exponential marginal models. In Section 3 simulation studies to assess the models estimates are exposed. Section 4 presents the results for the real dataset application of prostate cancer survival time. Lastly, in Section 5, final remarks are stated, as well as possible extensions of the present work.

2 Methodology

In this Section, we will present the likelihood formulation to model survival data with cure rate and dependent censoring. The approach allows a correlation between time to event of interest and censoring time in presence of *cured* individuals. In order to that promotion time cure rate model will be used. Copula functions are employed to build the joint probability distributions of time to event of interest and dependent censoring. In this formulation cure rate and survival times may take covariates into account.

Finally, we build the likelihood function for the cure rate model for dependent censoring under the copula-based approach. The proposed models are centered around Clayton and Plackett copulas with Weibull and piecewise exponential marginals distributions.

2.1 Cure rate model

A cure rate model is important when the population under study is composed of *susceptible* and *cured* individuals. All *susceptible* individuals can experience the event of interest if there is no censoring. On the other hand, *cured* individuals will not experience the event, but may be censored (Cancho et al., 2011).

In the approach proposed in this paper the promotion time cure rate model is considered. Following the formulation in Yakovlev and Tsodikov (1996) and the explanation in Chen et al. (1999) for this model suppose that for an individual of the study population, N denotes its number of carcinogenic cells left active after the initial treatment. Assume that N has a Poisson distribution with mean θ. Also let

R_j denote the random time for the j-th carcinogenic cell of this individual to produce a detectable cancer size. The variables R_j, j =1,2,... are assumed to be independent and identically distributed not related to N. The time of cancer relapse can be defined by the random variable $T = \min(R_0, R_1, ..., R_N)$, where $P(R_0 = \infty) = 1$. The survival function for T, and hence the survival function for the population, is given by $S_{pop}(t) = P(N = 0) + P(R_1 > t, ..., R_N > t|N \geq 1) \times P(N \geq 1)$, because of $N \sim Poisson(\theta)$ follows that the survival function for the population in time t is $S_{pop}(t) = \exp(-\theta F(t))$ for $t > 0$, which leads to

$$f_{pop}(t) = \theta f(t) \exp(-\theta F(t)) \ \ , h_{pop}(t) = \theta f(t) \ \text{ and } \ H_{pop}(t) = \theta F(t) \ \ , \qquad (1)$$

where $f_{pop}(t)$, $h_{pop}(t)$ and $H_{pop}(t)$ denote, respectively, the probability density function, the hazard function and the cumulative hazard function of the population under study in time t for $t > 0$, while $f(t)$ and $F(t)$ denote, respectively, the probability density function and the cumulative distribution function in time t for $t > 0$. The $f_{pop}(t)$ is said an improper density function because does not integrate to 1 as $\lim_{t \to \infty} S_{pop}(t) > 0$. Actually $\lim_{t \to \infty} S_{pop}(t) = \exp(-\theta)$, which means that the survival function for the population plateaus on $\exp(-\theta)$, this represents the proportion of individuals that will never experience the event of interest.

To formulate the cure rate model for dependent censoring we assume Y an observable random survival time variable defined as $Y = \min(T, C, A)$ where T, C and A are random positive variables, T is a random variable that denotes the time to event of interest of the *susceptible* population, C is a random variable that denotes the dependent censoring time and A is a random variable that denotes the independent administrative censoring time. Meanwhile y_i, t_i, c_i and a_i, $i = 1, 2, ..., n$ denote the i-th observation on Y, T, C and A respectively.

In real data however, only y_i is observable, along side with two indicator variables, $\delta_i = I\{y_i = t_i\}$ takes value 1 when the i-th individual experiences the event of interest and 0 otherwise. And $\rho_i = 1 - I\{y_i = ca_i\}$ takes value 1 when the i-th individual experiences either the event of interest or the dependent censoring and 0 otherwise.

The distributions of T and C can take into account covariates. Let $\mathbf{x}^T$ be a $n \times p$ matrix of p covariates associated with the time to event of interest distribution T and $\mathbf{x}^C$ a $n \times q$ vector of q covariates associated with the dependent censoring time distribution C. By Cox's proportional hazards model Cox (1972) the covariates take effect on the hazard function, where $\boldsymbol{\beta}^T$ is a vector $p \times 1$ of regression coefficients associated with $\mathbf{x}^T$ and $\boldsymbol{\beta}^C$, a $q \times 1$ vector of regression coefficients associated with $\mathbf{x}^C$.

The cure rate considered in the time to event of interest distribution is given by $p_0 = \exp(-\theta)$. It can also depend on covariates, where $\boldsymbol{\beta}$, a $s \times 1$ vector of regression coefficients associated with $\mathbf{x}$ a $n \times s$ matrix of s covariates. So, the populational hazard functions for the i-th individual $(i = 1, .., n)$ with cure rate is

$$h^T_{pop}(y_i|\mathbf{x}^T_i, \mathbf{x}_i) = h^T_0(y_i|\boldsymbol{\psi}^T) \exp\{\mathbf{x}_i\boldsymbol{\beta} + \mathbf{x}^T_i\boldsymbol{\beta}^T - H^T_0(y_i|\boldsymbol{\psi}^T) \exp\{\mathbf{x}^T_i\boldsymbol{\beta}^T\}\}, \qquad (2)$$

where $h_0^T(.|\boldsymbol{\psi}^T)$ and $H_0^T(.|\boldsymbol{\psi}^T)$ denote, respectively, the baseline hazard function and the cumulative baseline hazard function of the time to event of interest, $\boldsymbol{\psi}^T$ is a parameter vector that specifies the baseline hazard function of T. Whereas the hazard function of the dependent censoring for i-th individual $(i = 1, .., n)$ is

$$h^C(y_i|\mathbf{x}_i^C) = h_0^C(y_i|\boldsymbol{\psi}^C)\exp\{\mathbf{x}_i^C\boldsymbol{\beta}^C\}, \tag{3}$$

where $h_0^C(y|\boldsymbol{\psi}^C)$ denotes the baseline hazard function for the censoring time C and $\boldsymbol{\psi}^C$ denotes the parameter set associated with the baseline hazard function of C.

To simplify notation let $\boldsymbol{\theta}^T = (\boldsymbol{\psi}^T, \boldsymbol{\beta}^T, \boldsymbol{\beta})$, $\boldsymbol{\theta}^C = (\boldsymbol{\psi}^C, \boldsymbol{\beta}^C)$, $\mathbf{d}^T = (\mathbf{x}^T, \mathbf{x})$ and $\mathbf{d}^C = (\mathbf{x}^C)$.

Using equations (2) and (3) the probability density and cumulative distribution functions can be obtained. They are as follow:

$$\begin{aligned} f_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}^T) = h_0^T(y_i|\boldsymbol{\psi}^T)\exp\Bigg[\exp\Bigg\{\mathbf{x}_i\boldsymbol{\beta} - H_0(y_i|\boldsymbol{\psi}^T)e^{\mathbf{x}_i^T\boldsymbol{\beta}^T}\Bigg\} - \\ H_0^T(y_i|\boldsymbol{\psi}^T)e^{\mathbf{x}_i^T\boldsymbol{\beta}^T} - e^{\mathbf{x}_i\boldsymbol{\beta}} + \mathbf{x}_i\boldsymbol{\beta} + \mathbf{x}_i^T\boldsymbol{\beta}^T\Bigg], \end{aligned} \tag{4}$$

$$F_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}^T) = 1 - \exp\Bigg[\exp\Bigg\{\mathbf{x}_i\boldsymbol{\beta} - H_0^T(y_i|\boldsymbol{\psi}^T)e^{\mathbf{x}_i^T\boldsymbol{\beta}^T}\Bigg\} - e^{\mathbf{x}_i\boldsymbol{\beta}}\Bigg], \tag{5}$$

$$f^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}^C) = h_0^C(y_i|\boldsymbol{\psi}^C)\exp\Bigg\{\mathbf{x}_i^C\boldsymbol{\beta}^C - H_0^C(y_i|\boldsymbol{\psi}^C)e^{\mathbf{x}_i^C\boldsymbol{\beta}^C}\Bigg\}, \tag{6}$$

$$F^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}^C) = 1 - \exp\{-H_0^C(y_i|\boldsymbol{\psi}^C)e^{\mathbf{x}_i^C\boldsymbol{\beta}^C}\}. \tag{7}$$

Assuming independence between time to event of interest and the censoring time we have $P(T > y_i, C > y_i) = P(T > y_i) \times P(C > y_i)$, which unfolds in

$$\lim_{\Delta y_i \longrightarrow 0} P(T \in (y_i, y_i + \Delta y_i]), C > y_i) = f^T(y_i)S^C(y_i), \tag{8}$$

$$\lim_{\Delta y_i \longrightarrow 0} P(C \in (y_i, y_i + \Delta y_i]), T > y_i) = f^C(y_i)S^T(y_i). \tag{9}$$

However when the assumption of independence between time to event of interest T and censoring time C is not true, then $P(T > y_i, C > y_i) \neq P(T > y_i) \times P(C > y_i)$. Therefore, it is necessary to consider the joint distribution of event of interest and censoring times. So, the likelihood function is given by

$$\begin{aligned} L = \prod_{i=1}^{n}[\lim_{\Delta y_i \longrightarrow 0} P(T \in (y_i, y_i + \Delta y_i], C > y_i)]^{\delta_i\rho_i} \times \\ [\lim_{\Delta y_i \longrightarrow 0} P(C \in (y_i, y_i + \Delta y_i], T > y_i)]^{\rho_i(1-\delta_i)} \times \\ [P(T > y_i, C > y_i)]^{1-\rho_i}. \end{aligned} \tag{10}$$

Defining $P(T > y_i, C > y_i)$ without independence between T and C is not easy and requires more knowledge about the joint probability distribution. In this paper, we will obtain the joint probability distribution through the copula functions.

2.2 Copula functions

A copula function is a multivariate distribution with uniform marginals in the unit square (Nelsen, 2007). According to Sklar theorem (Sklar, 1959), be $W = (W_1, W_2, \cdots, W_k)$ a random vector and $\mathcal{C}$: $\mathbb{R}^k \rightarrow \mathbb{R}$ a copula function, $F(w_1, w_2, \cdots, w_k)$ a joint probability distribution function and $F^{W_i}(w_i)$, $i = 1, \cdots, k$ its marginals distributions, then there is a copula $\mathcal{C}$ such that $F(w_1, w_2, \cdots, w_k) = \mathcal{C}(F^{W_1}(w_1), F^{W_2}(w_2), \cdots F^{W_k}(w_k))$ for all $w_i \in [-\infty, \infty]$, $i = 1, \cdots, k$. In the 2-dimensional case in the survival context $\mathcal{C}_\nu(u, v)$ is a copula, where $(u, v) = (F^T(y_i), F^C(y_i))$ and ν is the copula parameters that outlines the dependency between the random variables T and C. One copula function example is the product copula $\mathcal{C}_\nu(u, v) = uv$ of independent time to event of interest and censoring time.

In survival analysis is also useful to define the joint survival function $\overline{\mathcal{C}}_\nu(u, v) = 1 - u - v + \mathcal{C}_\nu(u, v)$. Assuming $u(t) = P(T \leq t)$ and $v(c) = P(C \leq c)$ it is possible to build a model that allows correlation between times to event of interest T and dependent censoring time C through the copula parameter using the marginal distribution functions of T and C. The joint probabilities are expressed as

$$P(T > y_i, C > y_i) = \overline{\mathcal{C}}_\nu(u(y_i), v(y_i)), \tag{11}$$

$$\lim_{\Delta y_i \longrightarrow 0} P(T \in (y_i, y_i + \Delta y_i], C > y_i) = \frac{\partial \overline{\mathcal{C}}_\nu(u(t), v(c))}{\partial t}\Big|_{(t,c)=(y_i,y_i)}, \tag{12}$$

$$\lim_{\Delta y_i \longrightarrow 0} P(C \in (y_i, y_i + \Delta y], T > y_i) = \frac{\partial \overline{\mathcal{C}}_\nu(u(t), v(c))}{\partial c}\Big|_{(t,c)=(y_i,y_i)}. \tag{13}$$

There are many copulas functions to be chosen. We selected the Clayton copula and Plackett copula to study in this paper.Their joint survival function $\overline{\mathcal{C}}_\nu(u, v)$ are given by, respectively,

$$\overline{\mathcal{C}}_\nu(u(t), v(c)) = 1 - u(t) - v(c) + [u(t)^{-\nu} + v(c)^{-\nu} - 1]^{-\nu^{-1}} \quad \text{and} \tag{14}$$

$$\begin{aligned}\overline{\mathcal{C}}_\nu(u(t), v(c)) = 1 - u(t) - v(c) + \frac{1 + (\nu - 1)(u(t) + v(c))}{2(\nu - 1)} - \\ \frac{\sqrt{[1 + (\nu - 1)(u(t) + v(c))]^2 - 4u(t)v(c)\nu(\nu - 1)}}{2(\nu - 1)}.\end{aligned} \tag{15}$$

As seen in equations (12) and (13), the Clayton joint survival derivatives in relation to t and c are given by

$$\frac{\partial \overline{\mathcal{C}}_{\nu}(u(t), v(c))}{\partial t} = \big[u(t)^{-\nu-1}[u(t)^{-\nu} + v(c)^{-\nu} - 1]^{-\nu^{-1}-1} - 1\big]u(t)', \tag{16}$$

$$\frac{\partial \overline{\mathcal{C}}_{\nu}(u(t), v(c))}{\partial c} = \big[v(c)^{-\nu-1}[u(t)^{-\nu} + v(c)^{-\nu} - 1]^{-\nu^{-1}-1} - 1\big]v(c)'. \tag{17}$$

The Plackett joint survival function derivatives are given by:

$$\begin{aligned}\frac{\partial \overline{\mathcal{C}}_{\nu}(u(t), v(c))}{\partial t} = \big[-&\frac{[(\nu-1)(u(t)+v(c))+1] - 4(\nu-1)\nu v(c)}{2\sqrt{[(\nu-1)(u(t)+v(c))+1]^2 - 4(\nu-1)\nu v(c) u(t)}} + \\ &\frac{\nu-1}{2(\nu-1)} - 1\big]u(t)',\end{aligned} \tag{18}$$

$$\begin{aligned}\frac{\partial \overline{\mathcal{C}}_{\nu}(u(t), v(c))}{\partial c} = \big[-&\frac{[(\nu-1)(v(c)+u(t))+1] - 4(\nu-1)\nu u(t)}{2\sqrt{[(\nu-1)(v(c)+u(t))+1]^2 - 4(\nu-1)\nu u(t) v(c)}} + \\ &\frac{\nu-1}{2(\nu-1)} - 1\big]v(c)'.\end{aligned} \tag{19}$$

Now, we can formulate the cure rate model for dependent censoring likelihood function as in equation (10) using copulas. To that use equations (11), (12) and (13) in (10). The likelihood function is given by

$$\begin{aligned}L = \prod_{i=1}^{n} &\left[\frac{\partial \overline{\mathcal{C}}_{\nu}(u(t), v(c))}{\partial t}\Big|_{(t,c)=(y,y)}\right]^{\delta_i \rho_i} \times \\ &\left[\frac{\partial \overline{\mathcal{C}}_{\nu}(u(t), v(c))}{\partial t}\Big|_{(t,c)=(y,y)}\right]^{\rho_i(1-\delta_i)} \times \\ &\left[\overline{\mathcal{C}}_{\nu}(u(t), v(c))\Big|_{(t,c)=(y,y)}\right]^{(1-\rho_i)}\end{aligned} \tag{20}$$

The Clayton copula parameter ν is related to Kendall's τ correlation coefficient, whereas the Plackett copula parameter ν is related to Spearman's correlation coefficient ρ. The relations are as follow

$$\tau_s(\text{Clayton's } \ \nu) = \frac{\nu}{\nu+2} \quad \text{and} \quad \rho_s(\text{Plackett's } \ \nu) = \frac{\nu+1}{\nu-1} - \frac{2\nu\ln(\nu)}{(\nu-1)^2}. \tag{21}$$

There is no closed form to find the Spearman's correlation coefficient ρ under the Clayton copula, neither there is a closed form to find the Kendall's correlation coefficient τ under the Plackett copula. In order to find copula parameters for the Clayton and Plackett that provide the same correlation, it is necessary to use computational procedures.

Based on the marginal formulations about T and C and the copula functions we can express the likelihood function of the cure rate model for dependent censoring, given by

$$\begin{aligned} &L(\boldsymbol{\theta}^T, \boldsymbol{\theta}^C | \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\rho}, \nu) = \\ &\prod_{i=1}^{n} \left[f_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T) \frac{\partial \overline{C}_\nu (F_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T), F^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C))}{\partial F_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T)} \right]^{\delta_i \rho_i} \times \\ &\left[f^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C) \frac{\partial \overline{C}_\nu (F_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T), F^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C))}{\partial F^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C)} \right]^{(1-\delta_i)\rho_i} \times \\ &\left[\overline{C}_\nu (F_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T), F^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C)) \right]^{(1-\rho_i)}. \end{aligned} \tag{22}$$

Where $\mathbf{y}$ and the indicators variable $\boldsymbol{\delta}$ and $\boldsymbol{\rho}$ compose the set of observed data. $\boldsymbol{\theta}^T$ and $\boldsymbol{\theta}^C$ denote the parameter vectors associated with the T and C respectively, $\mathbf{d}^T$ and $\mathbf{d}^C$ denote the set of covariates.

To obtain maximum likelihood estimates of $\boldsymbol{\theta}^T$ and $\boldsymbol{\theta}^C$ solve the gradient vector of the likelihood function to zero, or more easily the gradient vector of the logarithm of the likelihood function. As there is not an analytical solution to these partial derivatives an optimization algorithm is required. Standard errors can be obtained from the square roots of the negative Hessian's inverse principal diagonal $SE = \sqrt{-\text{diag}(\mathcal{H}^{-1})}$.

The likelihood function presented in equation (22) besides the copula function also needs marginal distributions specification for T and C. In survival analysis, appropriate distributions need to be chosen to model positive data. There are many options of distributions, each one has benefits and limitations. In this paper, we will use the Weibull and piecewise exponential distributions to model event of interest and dependent censoring times.

2.3 Marginal Weibull Model

The Weibull distribution is a common choice in survival analysis, with a monotonic hazard function and two parameters it has the exponential distribution as a particular case (Johnson et al., 1995). Assuming that the time to event of interest $T \sim W(\alpha_T, \lambda_T)$, $\alpha_T, \lambda_T > 0$ and dependent censoring time $C \sim W(\alpha_C, \lambda_C)$, with $\alpha_C, \lambda_C > 0$; α^T and α^C are shape parameters, λ^T and λ^C are the scale parameters. For this model we have the following notations $\boldsymbol{\psi}_T = (\alpha_T, \lambda_T)$ and $\boldsymbol{\psi}_C = (\alpha_C, \lambda_C)$. The baseline hazard functions and cumulative baselines hazard functions are

$$h_0^T(y|\boldsymbol{\psi}^T) = \alpha^T\lambda^T y^{\alpha^T-1} \text{ and } h_0^C(y|\psi^C) = \alpha^C\lambda^C y^{\alpha^C-1}, \tag{23}$$

$$H_0^T(y|\boldsymbol{\psi}^T) = \lambda^T y^{\alpha^T} \text{ and } H_0^C(y|\psi^C) = \lambda^C y^{\alpha^C}. \tag{24}$$

To obtain the likelihood function of the cure rate model for dependent censoring under copulas expressed in equation (22), with Weibull distributions for T and C we need $f_{pop}^T(y|\boldsymbol{\theta}^T, \mathbf{d}^T)$, $F_{pop}^T(y|\boldsymbol{\theta}^T, \mathbf{d}^T)$, $f^C(y|\boldsymbol{\theta}^C, \mathbf{d}^C)$, $F^C(y|\boldsymbol{\theta}^T, \mathbf{d}^T)$. To define these functions use equations (23) and (24) in equations (4),(5),(6),(7), resulting in the following functions

$$\begin{aligned} f_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T) = \alpha_T\lambda_T {y_i}^{\alpha_T-1}\exp\Bigg[\exp\Bigg\{\mathbf{x}_i\boldsymbol{\beta} - \lambda_T {y_i}^{\alpha_T} e^{\mathbf{x}_i^T\boldsymbol{\beta}^T}\Bigg\} - \\ \lambda_T {y_i}^{\alpha_T} e^{\mathbf{x}_i^T\boldsymbol{\beta}^T} - e^{\mathbf{x}_i\boldsymbol{\beta}} + \mathbf{x}_i\boldsymbol{\beta} + \mathbf{x}_i^T\boldsymbol{\beta}^T\Bigg], \end{aligned} \tag{25}$$

$$F_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T) = 1 - \exp\Bigg[\exp\Bigg\{\mathbf{x}_i\boldsymbol{\beta} - \lambda_T {y_i}^{\alpha_T} e^{\mathbf{x}_i^T\boldsymbol{\beta}^T}\Bigg\} - e^{\mathbf{x}_i\boldsymbol{\beta}}\Bigg], \tag{26}$$

$$f^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C) = \alpha^C\lambda^C {y_i}^{\alpha^C-1}\exp\{\mathbf{x}_i^C\boldsymbol{\beta}^C - \lambda^C {y_i}^{\alpha^C} e^{\mathbf{x}_i^C\boldsymbol{\beta}^C}\}, \tag{27}$$

$$F^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C) = 1 - \exp\{-\lambda^C {y_i}^{\alpha^C} e^{\mathbf{x}_i^C\boldsymbol{\beta}^C}\}. \tag{28}$$

2.4 Marginal Piecewise Exponential Model

The piecewise exponential distribution is a semiparametric distribution, it makes no limitation on the shape of the hazard function and this is why it has become a popular option in survival analysis (Emura and Michimae, 2017). The piecewise exponential distribution has a step hazard function, constant over each interval of a time grid.

Assume a time grid for time to event of interest T as $\gamma^T = \{m_0^T, m_1^T, \cdots, m_b^T\}$, with $m_0^T = 0$ and $m_b^T = \infty$, where $m_0^T < m_1^T < \cdots < m_b^T$, creating a set of b disjoint intervals $I_k^T = (m_{k-1}^T, m_k^T]$ for $k = 1, \ldots, b$. Assume a similar grid for censoring time C as $\gamma^C = \{m_0^C, m_1^C, \ldots, m_d^C\}$ with $m_0^C = 0$ and $m_d^C = \infty$, where $m_0^C < m_1^C < \cdots < m_d^C$, creating a set of d disjoint intervals $I_j^C = (m_{j-1}^C, m_j^C]$ for $j = 1, \ldots, d$.

Consider $\boldsymbol{\psi}^T$ to denote the vector of λ_k^T for $k = 1, \ldots, b$ that specifies the constant hazard function in each interval defined by the grid γ^T . And $\boldsymbol{\psi}^C$ to denote the vector of λ_j^C for $j = 1, \ldots, d$ that specifies the constant hazard function in each interval defined by the grid γ^C. The baselines hazard function for the piecewise exponential distribution are

$$
\begin{aligned}
h_0^T(y|\boldsymbol{\psi}^T) &= \sum_{k=1}^{b} \lambda_k^T I(m_{k-1}^T \leq y < m_k^T) \\
\text{and} \quad h_0^C(y|\boldsymbol{\psi}^C) &= \sum_{j=1}^{d} \lambda_j^C I(m_{j-1}^C \leq y < m_j^C) \ ,
\end{aligned} \tag{29}
$$

$$
\begin{aligned}
H_0^T(y|\boldsymbol{\psi}_T) &= \sum_{k=1}^{b} \lambda_k^T (\min\{y, m_k^T\} - \min\{y, m_{k-1}^T\}) \\
\text{and} \quad H_0^C(y|\boldsymbol{\psi}_C) &= \sum_{j=1}^{d} \lambda_j^C (\min\{y, m_j^C\} - \min\{y, m_{j-1}^C\}) \ .
\end{aligned} \tag{30}
$$

To obtain the likelihood function of the cure rate model for dependent censoring under copulas expressed in equation (22) with piecewise exponential distribution for T and C we need $f_{pop}^T(y|\boldsymbol{\theta}^T, \mathbf{d}^T)$, $F_{pop}^T(y|\boldsymbol{\theta}^T, \mathbf{d}^T)$, $f^C(y|\boldsymbol{\theta}^C, \mathbf{d}^C)$, $F^C(y|\boldsymbol{\theta}^T, \mathbf{d}^T)$. To derive them use equations (29) and (30) in equations (4),(5),(6),(7) , resulting in the following functions

$$
\begin{aligned}
f_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T) &= \left[\sum_{k=1}^{b} \lambda_k^T I(m_{k-1}^T \leq y_i < m_k^T)\right] \times \\
\exp\Bigg\{ \exp\Bigg\{ \mathbf{x}_i\boldsymbol{\beta} &- \left[\sum_{k=1}^{b} \lambda_k^T (\min\{y_i, m_k^T\} - \min\{y_i, m_{k-1}^T\})\right] e^{\mathbf{x}_i^T \boldsymbol{\beta}^T} \Bigg\} - \\
\left[\sum_{k=1}^{b} \lambda_k^T (\min\{y_i, m_k^T\} - \min\{y_i, m_{k-1}^T\})\right] & e^{\mathbf{x}_i^T \boldsymbol{\beta}^T} - e^{\mathbf{x}_i\boldsymbol{\beta}} + \mathbf{x}_i\boldsymbol{\beta} + \mathbf{x}_i^T\boldsymbol{\beta}^T \Bigg\},
\end{aligned} \tag{31}
$$

$$
\begin{aligned}
F_{pop}^T(y_i|\boldsymbol{\theta}^T, \mathbf{d}_i^T) = 1 - \exp\Bigg[\exp\Bigg\{ \mathbf{x}_i\boldsymbol{\beta} - \Bigg[\sum_{k=1}^{b} \lambda_k^T (\min\{y_i, m_k^T\} - \\
\min\{y_i, m_{k-1}^T\}) \Bigg]^{\mathbf{x}_i^T \boldsymbol{\beta}^T} \Bigg\} - e^{\mathbf{x}_i\boldsymbol{\beta}} \Bigg],
\end{aligned} \tag{32}
$$

$$
\begin{aligned}
f^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C) &= \left[\sum_{j=1}^{d} \lambda_j^C I(m_{k-1}^C \leq y_i < m_j^C)\right] \times \\
\exp\Bigg\{ \mathbf{x}_i^C\boldsymbol{\beta}^C &- \left[\sum_{j=1}^{d} \lambda_j^C (\min\{y_i, m_j^C\} - \min\{y_i, m_{k-1}^C\})\right] e^{\mathbf{x}_i^C \boldsymbol{\beta}^C} \Bigg\},
\end{aligned} \tag{33}
$$

$$F^C(y_i|\boldsymbol{\theta}^C, \mathbf{d}_i^C) = 1 - \exp\left\{\left[\sum_{j=1}^{d} \lambda_j^C(\min\{y_i, m_j^C\} - \min\{y_i, m_{k-1}^C\})\right] e^{\mathbf{x}_i^C \boldsymbol{\beta}^C}\right\}. \quad (34)$$

3 Simulation Study

To assess the proposed cure rate model for dependent censoring under copula approach we conduct a Monte Carlo simulation study. The implementation was done in software R version 4.0.5. We generated 500 data sets with $n = 500$ assuming Clayton copula with $\nu = 3$, producing moderate correlation, Kendall's τ of $\sim 25\%$. For the time distributions we considered the Weibull model with parameters $\alpha^T = 1.5$, $\lambda^T = 0.2$, $\alpha^C = 1.5$ and $\lambda^C = 0.15$. The administrative censoring A time was considered uniform between 0 and 20. For the covariates related to T, C and the cure rate of T, we generated X_1^T ,X_1^C and X_1 from $Ber(0.5)$ and X_2^T, X_2^C and X_2 from $N(0;1)$. The regression coefficients were set as $\boldsymbol{\beta}^T = (1.2, -1.2)$, $\boldsymbol{\beta}^C = (-1.4, 1.4)$ and $\boldsymbol{\beta} = (-0.6, 0.7, 0.8)$.

To generate data from cure rate model for dependent censoring under Clayton copula with Weibull marginals, set $t_i = {F_{pop}^T}^{-1}(u_i|\theta^T, \mathbf{d}_i^T)$, $c_i = F^{C^{-1}}(v_i|\theta^C, \mathbf{d}_i^C)$ and $a_i \sim U[0; 20]$, Because of the dependency context u_i and v_i are correlated.

In order to generate correlated u_i and v_i, we use the Clayton copula. First, draw u_i from $U[0; 1]$, then draw v_i from $P(V|U = u_i)$. $v_i = P(V \leq w_i|U = u_i)^{-1}$ where w_i is $U[0; 1]$. For Clayton copula we have:

- Draw $u_i \sim U[0; 1]$
- Draw $w_i \sim U[0; 1]$
- Take $v_i = [u_i^{-\nu}(w_i^{-\nu(\nu+1)^{-1}} - 1) + 1]^{-1/\nu}$.

Set $y_i = \min(t_i, c_i, a_i)$ to define the observed survival time and let $\delta_i = I\{y_i = t_i\}$ and $\rho_i = 1 - I\{y_i = a_i\}$ be the indicator variables. To obtain t_i and c_i use respectively,

$$t_i = \left[-\ln\left(1 + \frac{\ln(1 - u_i)}{\exp\{\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{1,i} + \boldsymbol{\beta}_2 X_{2,i}\}}\right) \times \frac{1}{\lambda_T \exp\{\boldsymbol{\beta}_1^T X_{1,i}^T + \boldsymbol{\beta}_2^T X_{2,i}^T\}}\right]^{\alpha^{T-1}} \quad (35)$$

$$\text{and} \quad c_i = \left[\frac{-\ln(1 - v_i)}{\lambda_C \exp\{\boldsymbol{\beta}_1^C X_{1,i}^C + \boldsymbol{\beta}_2^C X_{2,i}^C\}}\right]^{\alpha^{C-1}}. \quad (36)$$

The Clayton's copula parameter ν is related to Kendall's τ as presented by equation (21).This Kendall's τ however expresses the correlation value between U and V, not between T and C. The data was generated setting $\nu = 3$, which denotes

Kendall's $\tau = 0.60$ between U and V, while correlation between T and C is Kendall's $\tau = 0.2573$.

The above specifications generate samples with approximately 38.3% event outcome, 39, 2% dependent censoring and 22.5% administrative censoring. Each one of the 500 datasets is fit by the proposed models: Clayton copula cure model for dependent censoring with Weibull marginals (M1) (generator model) ; Clayton copula cure model for dependent censoring with piecewise exponential marginals (M2); Plackett copula cure model for dependent censoring with Weibull marginals (M3) and Plackett copula cure model for dependent censoring with piecewise exponential marginals (M4). We fit each model for four different dependency scenarios. We do this to evaluate the model parameters estimations and to compare the different dependency scenarios. Table 1 shows the copulas parameters set in each scenario.

Table 1 - Correlation Scenarios

Scenario	**Clayton's ν**	**Plackett's ν**	**Kendall's τ**
Indep	0.0001	1.0001	∼0.0005
Under	1	5	∼0.15
Correct	3	50	∼0.25
Over	8	300	∼0.30

The Akaike information criterion and percentage of lowest AIC were used to determine the adequate number of intervals (see Table 2) for the piecewise exponential marginal models in M2 and M4. Ten datasets were adjusted by M1 assuming different number of intervals.The adjustments assuming 5 intervals provided good mean AIC and presented the highest percentage of lowest AIC. The intervals cut points were defined using the observed times, allocating the same number of observations in each interval.

Table 2 - Number of intervals for piecewise exponential marginal

Intervals	***Weibull***	**2**	**3**	**4**	**5**	**6**	**7**	**10**	**15**
$\overline{\textbf{AIC}}$	*1490*	1545	1505	1497	1497	1579	1681	2111	2374
% Lowest		10%	20%	20%	40%	10%	0%	0%	0%

$\overline{\textbf{AIC}}$ is the mean AIC of the ten datasets's AIC.; **% Lowest** is the lowest AIC percentage of interval quantity

Estimates of the parameters were found through Maximum likelihood using a quasi-Newton method. This optimizer is available in the R software under BFGS method in the *optim* function that uses the Shanno (1970) approach. To start the gradient guided quasi-Newton we first conducted one run of Nelder and Mead optimizer with initial values of 0.01 for all parameters and used its estimates as initial values for the BFGS optimizer. To obtain the Hessian matrix we used the Richardson extrapolation method to approximate derivatives implemented under the function *hessian* in *numDeriv* package in the R software .

So, each one of the 500 datasets are fitted by four Models (M1, M2, M3, M4) and four dependency scenarios (*Indep*, *Under*, *correct*, *Over*). The Tables 3, 4, 5, 6 show the summarized estimates obtained from the fitted models. In order to evaluate the proposed model's estimates for different correlation scenarios the Tables inform the mean estimates (**Est**) of each regression coefficient $\boldsymbol{\beta}^T$,$\boldsymbol{\beta}^C$ and $\boldsymbol{\beta}$, comparing them to their true values **Real** through the mean relative bias (**%Bias**). **CP** denotes the Coverage Probability, **SE** is the mean standard error and **SD** is the standard deviation of the Monte Carlo replications. Furthermore we calculated the mean of the Bayesian information criterion (BIC), Hannan - Quinn information criterion (HQ) and Akaike information criterion (AIC).

Table 3 presents the summary of M1 estimates. M1 is the data generator model, and so is expected to present good estimates specially for the *correct* scenario. Assuming independence M1 produces very poor estimates. Relative bias of β_1^C, β_0^C and β_0 are over 30% and the coverage probabilities of these parameters are low, specially β_2^C whose coverage probability is 0.012 (Table 3). The results of M1 in the *Indep* scenario reiterate the importance of modeling the dependence. All parameters estimates present positive relative bias, in other words the model overestimate parameters. When we look to the M1 model in the *Under* scenario and compare it to *Indep*, relative bias is smaller then those of *Indep* for all parameters. Coverage probability is higher for all parameters, but still far from the nominal 0.95 (Table 3).

Under the *correct* scenario M1 generated good estimates, highest relative bias is β_2^T's of only 1.78%. Coverage probability of all parameters also reach close to nominal level. The *Over* scenario however produced poor estimates with large negative relative bias, β_0 the cure rate's intercept presents the largest relative bias (-18.32%), all parameters also present smaller than ideal coverage probability when compared to the nominal value (Table 3).

In Table 4 we present the summary of the M3 models estimates. The results are very similar to those seen in M1 model estimates. The *Indep* scenario produces large (from 10.7% to 35.7%) positive bias estimates with low (0.012 - 0.88) coverage probability. Again, β_2^C present worst metrics, 0.012 coverage probability and 31.7% relative bias. The *Under* scenario shows improvements for all parameters estimates when compared to *Indep*, the bias is reduced and coverage probability increases. The *correct* scenario produces good estimates, this is important because M3 is not the data generator model. M3 model uses the Plackett copula to adjust for dependency instead the Clayton copula. The *correct* scenario presented good coverage probability for all parameters, and all parameters except β_2 have positive small relative bias (from 0.5% to 1.9%) (Table 4). The *Over* scenario produced mostly negative biased estimates, only β_0 presented positive relative bias (1.087%), this bias is actually smaller then that seen in the *correct* scenario (1.905%). All other parameters, besides β_0, present relative bias smaller than -5% (Table 4).

Table 5 presents the summary of the M2 model estimates. The M2 model uses Clayton copula to adjust the dependency and assumes piecewise exponential marginals. Once again, the model in the *Indep* scenario produces estimates with large positive relative bias (from 7.9% to 32.9%) and low coverage probability (from 0.91 to 0.04). The parameter β_0 presents the largest relative bias (32.9%) and again, β_2^C has an extremely low coverage probability (0.04) whereas β_1^T however presents much better (0.91) . The M2 model in *Under* scenario shows improvements for all parameters estimates in relation to de *Indep* scenario, relatives bias lowers and coverage probability increases. The parameters β_1^T, β_2^T and β_1 present very high coverage probability (0.94, 0.93, 0.94). For the *correct* scenario all parameters present negative relative bias (from -2.5% to -0.6%), although close to zero, this diverges from the other models that had mostly small positive relative bias (See Tables 3 and 4) . Coverage probability value is close to the nominal level, for all parameters. However β_1^T, β_2^T and β_1 present reduced coverage probability in comparison with the *Under* scenario (Table 5). The M2 model in the *Over* scenario produces large negative biased estimates (from -23.5% to -15.4%)) with low coverage probability (from 0.12 to 0.61).The worst behavior is for the β_0 parameter that presents -23.5% relative bias. These parameters presented the worst relative bias in all scenarios of M2 model(Table 5). Comparing the *Over* scenarios, M2 model has the worst performance between M1 and M3 in terms of relative bias for all parameters. Even so, M2 model in the *Over* scenario is better than the *Indep* scenario for β_1^C,β_2^C and β_0 (Table 5).

Table 6 presents the summary of M4 model estimates. The M4 model uses Plackett copula to adjust the dependency and assumes piecewise exponential marginals, this is important because M4 model differs completely from M1 model(data generator model), M1 models uses Clayton copula to adjust the dependency and assumes Weibull marginals. The M4 model in the *Indep* scenario produces estimates with large positive bias (from 7.8% to 32.8%). The parameter β_0 with 32.8% has the highest relative bias. the parameter β_1^T presents a high coverage probability (0.912). The *Under* scenario presents improved estimates in comparison to *Indep* scenario present improved estimates because all the parameters reduced bias and increased coverage probability. Relative bias is still large (from 6% to 16%). Parameters β_1^T, β_2^T, β_0 and β_1 present high coverage probability (0.92, 0.9, 0.92) (Table 6). The M4 model in the *correct* scenario presents negative close to zero bias (from -3.1% to -1.8%) for all parameters and good coverage probability (from 0.92 to 0.95). Parameter β_1^T presents reduced coverage probability in comparison to the *Under* scenario, but still is close to the nominal level. The *Over* scenario present all the parameter's estimates with negative relative bias (from -11.2% to -4.8%) and loss coverage probability, compared to *Under* and *correct* scenarios (Table 6).

Analyzing the simulations results as a whole we can see that the mean standard errors are close to the standard deviation in all models and all scenarios. The coverage probability reaches the nominal level when correctly setting the dependency.

Estimates of all four models show increased bias as the dependency misspecification gets worse. In our simulation for *Indep* scenario all models presented large positive relative bias, while for the *Over* scenario all models presented large negative bias. Simulations show that ,to a certain degree, it is better to misspecify the dependency than to assume independence.

Comparing models that used Plackett copula and models that used Clayton copula no differences draws attention. Models are much more sensitive to the copula parameter than the copula choice. Comparing the models that assume Weibull marginals and the models that assume piecewise exponential, showed that the piecewise exponential models are able to fit the data generated from the Weibull model as expected. Even though the piecewise exponential models presented negative relative bias further from zero than Weibull, in the *Over* scenario the piecewise exponential model presented higher coverage probability for some parameters.

Assessing all parameters of each scenario, the *Indep* scenario presented the worst estimates, followed by the *Over* scenario, then the *Under*. This indicates that for our simulation both misspecification of dependency scenarios are better then the *Indep* scenario. M3 model performed better in terms of bias than the M1 model for the *Over* scenario. All parameters present lower relative bias in the *Over* scenario of M3 model than in *Over* scenario of M1 model.

A broader simulation might elucidate more about the models behaviors, specially the comparison of the piecewise exponential and the Weibull model, in this simulation study we only generated data from the M1 model, it would be interesting to generate data from other models as well. Although the copula function did not matter much we know that the marginal model may, as the piecewise exponential model is able to fit non-monotonic hazard functions.

Table 3 - Simulations results for the proposed model M1:Clayton copula cure model for dependent censoring with Weibull marginals

Model		**Par**	**Real**	**Est**	**%Bias**	**CP**	**SE**	**SD**
M1		β_1^T	1.2	1.330	10.870	0.880	0.190	0.203
		β_2^T	- 1.2	- 1.329	10.785	0.774	0.110	0.110
Indep		β_1^C	- 1.4	- 1.850	32.110	0.236	0.168	0.174
		β_2^C	1.4	1.845	31.757	0.012	0.111	0.120
$\overline{\text{BIC}}$	1597.0	β_0	- 0.6	- 0.814	35.721	0.648	0.133	0.130
$\overline{\text{HQ}}$	1568.9	β_1	0.7	0.812	15.954	0.882	0.152	0.152
$\overline{\text{AIC}}$	1550.7	β_2	0.8	0.914	14.274	0.724	0.083	0.083
M1		β_1^T	1.2	1.306	8.843	0.898	0.177	0.183
		β_2^T	- 1.2	- 1.307	8.941	0.812	0.103	0.101
Under		β_1^C	- 1.4	- 1.624	15.993	0.684	0.142	0.141
		β_2^C	1.4	1.622	15.872	0.340	0.094	0.096
$\overline{\text{BIC}}$	1557.9	β_0	- 0.6	- 0.713	18.760	0.868	0.122	0.118
$\overline{\text{HQ}}$	1529.7	β_1	0.7	0.777	11.058	0.926	0.138	0.134
$\overline{\text{AIC}}$	1511.5	β_2	0.8	0.877	9.675	0.824	0.076	0.075
M1		β_1^T	1.2	1.216	1.340	0.962	0.155	0.159
		β_2^T	- 1.2	- 1.221	1.783	0.938	0.092	0.094
Correct		β_1^C	- 1.4	- 1.416	1.139	0.936	0.118	0.120
		β_2^C	1.4	1.416	1.178	0.936	0.081	0.085
$\overline{\text{BIC}}$	1536.6	β_0	- 0.6	- 0.611	1.762	0.956	0.108	0.109
$\overline{\text{HQ}}$	1508.4	β_1	0.7	0.712	1.770	0.960	0.117	0.115
$\overline{\text{AIC}}$	1490.2	β_2	0.8	0.806	0.780	0.946	0.068	0.070
M1		β_1^T	1.2	1.045	- 12.933	0.688	0.124	0.156
		β_2^T	- 1.2	- 1.058	- 11.829	0.538	0.078	0.102
Over		β_1^C	- 1.4	- 1.198	- 14.433	0.440	0.094	0.120
		β_2^C	1.4	1.199	- 14.359	0.230	0.070	0.089
$\overline{\text{BIC}}$	1565.9	β_0	- 0.6	- 0.490	- 18.321	0.710	0.093	0.111
$\overline{\text{HQ}}$	1537.7	β_1	0.7	0.607	- 13.327	0.754	0.089	0.112
$\overline{\text{AIC}}$	1519.5	β_2	0.8	0.689	- 13.921	0.500	0.056	0.075

Table 4 - Simulations results for the proposed model M3:Plackett copula cure model for dependent censoring with Weibull marginals

Model		**Par**	**Real**	**Est**	**%Bias**	**CP**	**SE**	**SD**
M3		β_1^T	1.2	1.330	10.870	0.880	0.190	0.203
		β_2^T	- 1.2	- 1.329	10.785	0.774	0.110	0.110
Indep		β_1^C	- 1.4	- 1.850	32.124	0.236	0.168	0.174
		β_2^C	1.4	1.845	31.770	0.012	0.111	0.120
$\overline{\text{BIC}}$	1597.1	β_0	- 0.6	- 0.814	35.736	0.648	0.133	0.130
$\overline{\text{HQ}}$	1568.9	β_1	0.7	0.812	15.957	0.882	0.152	0.152
$\overline{\text{AIC}}$	1550.7	β_2	0.8	0.914	14.277	0.724	0.083	0.083
M3		β_1^T	1.2	1.312	9.367	0.902	0.182	0.190
		β_2^T	- 1.2	- 1.315	9.587	0.818	0.107	0.104
Under		β_1^C	- 1.4	- 1.672	19.402	0.586	0.148	0.152
		β_2^C	1.4	1.673	19.498	0.216	0.101	0.104
$\overline{\text{BIC}}$	1567.9	β_0	- 0.6	- 0.703	17.246	0.890	0.125	0.121
$\overline{\text{HQ}}$	1539.8	β_1	0.7	0.785	12.079	0.924	0.140	0.137
$\overline{\text{AIC}}$	1521.6	β_2	0.8	0.885	10.584	0.822	0.078	0.078
M3		β_1^T	1.2	1.206	0.539	0.938	0.160	0.168
		β_2^T	- 1.2	- 1.215	1.260	0.958	0.096	0.096
Correct		β_1^C	- 1.4	- 1.414	1.015	0.936	0.120	0.128
		β_2^C	1.4	1.419	1.354	0.954	0.087	0.088
$\overline{\text{BIC}}$	1551.9	β_0	- 0.6	- 0.611	1.905	0.936	0.109	0.113
$\overline{\text{HQ}}$	1523.8	β_1	0.7	0.702	0.290	0.954	0.114	0.122
$\overline{\text{AIC}}$	1505.6	β_2	0.8	0.794	- 0.789	0.938	0.069	0.077
M3		β_1^T	1.2	1.118	- 6.836	0.862	0.141	0.170
		β_2^T	- 1.2	- 1.129	- 5.939	0.818	0.088	0.102
Over		β_1^C	- 1.4	- 1.293	- 7.670	0.744	0.107	0.133
		β_2^C	1.4	1.298	- 7.280	0.718	0.082	0.091
$\overline{\text{BIC}}$	1574.1	β_0	- 0.6	- 0.607	1.087	0.890	0.100	0.122
$\overline{\text{HQ}}$	1546.0	β_1	0.7	0.650	- 7.073	0.812	0.097	0.134
$\overline{\text{AIC}}$	1527.8	β_2	0.8	0.736	- 8.050	0.744	0.064	0.086

Table 5 - Simulations results for the proposed model M2: Clayton copula cure model for dependent censoring with Piecewise Exponential marginals

Model		**Par**	**Real**	**Est**	**%Bias**	**CP**	**SE**	**SD**
M2		β_1^T	1.2	1.302	8.523	0.912	0.191	0.201
		β_2^T	- 1.2	- 1.296	7.996	0.874	0.110	0.105
Indep		β_1^C	- 1.4	- 1.804	28.842	0.322	0.168	0.166
		β_2^C	1.4	1.786	27.571	0.040	0.107	0.108
$\overline{\text{BIC}}$	1645.8	β_0	- 0.6	- 0.797	32.839	0.696	0.134	0.129
$\overline{\text{HQ}}$	1602.3	β_1	0.7	0.805	15.014	0.896	0.152	0.150
$\overline{\text{AIC}}$	1574.2	β_2	0.8	0.905	13.162	0.756	0.082	0.081
M2		β_1^T	1.2	1.264	5.337	0.948	0.177	0.180
		β_2^T	- 1.2	- 1.260	5.037	0.934	0.101	0.096
Under		β_1^C	- 1.4	- 1.594	13.828	0.764	0.142	0.137
		β_2^C	1.4	1.586	13.264	0.500	0.093	0.090
$\overline{\text{BIC}}$	1606.1	β_0	- 0.6	- 0.688	14.682	0.922	0.123	0.116
$\overline{\text{HQ}}$	1562.5	β_1	0.7	0.764	9.203	0.946	0.137	0.131
$\overline{\text{AIC}}$	1534.4	β_2	0.8	0.861	7.664	0.896	0.075	0.073
M2		β_1^T	1.2	1.170	- 2.474	0.936	0.154	0.157
		β_2^T	- 1.2	- 1.172	- 2.322	0.928	0.090	0.089
Correct		β_1^C	- 1.4	- 1.380	- 1.393	0.946	0.118	0.118
		β_2^C	1.4	1.380	- 1.461	0.956	0.081	0.082
$\overline{\text{BIC}}$	1584.4	β_0	- 0.6	- 0.585	- 2.534	0.950	0.108	0.106
$\overline{\text{HQ}}$	1540.8	β_1	0.7	0.696	- 0.608	0.958	0.116	0.112
$\overline{\text{AIC}}$	1512.7	β_2	0.8	0.786	- 1.724	0.958	0.066	0.067
M2		β_1^T	1.2	1.005	- 16.245	0.608	0.123	0.155
		β_2^T	- 1.2	- 1.015	- 15.414	0.356	0.077	0.097
Over		β_1^C	- 1.4	- 1.151	- 17.810	0.284	0.094	0.118
		β_2^C	1.4	1.153	- 17.656	0.126	0.070	0.088
$\overline{\text{BIC}}$	1610.7	β_0	- 0.6	- 0.458	- 23.594	0.618	0.093	0.108
$\overline{\text{HQ}}$	1567.1	β_1	0.7	0.585	- 16.399	0.678	0.088	0.109
$\overline{\text{AIC}}$	1539.0	β_2	0.8	0.664	- 16.990	0.350	0.055	0.073

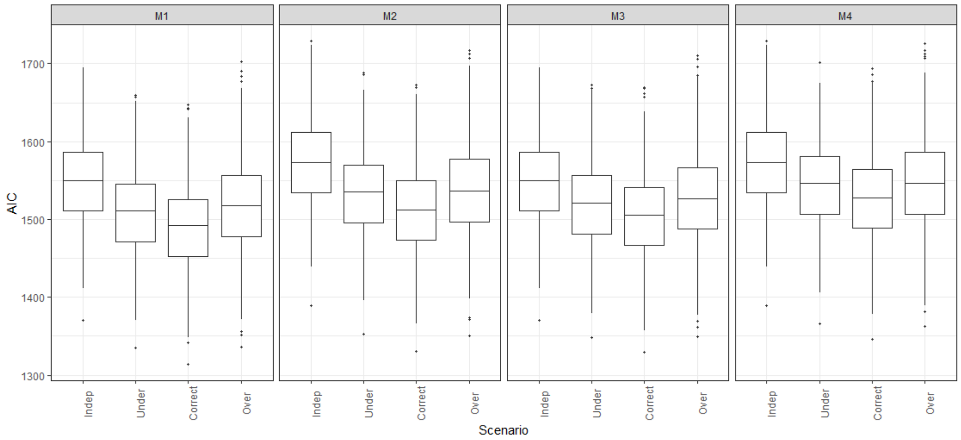
Table 6 - Simulations results for the proposed model M4: Plackett copula cure model for dependent censoring with Piecewise Exponential marginals

Model		**Par**	**Real**	**Est**	**%Bias**	**CP**	**SE**	**SD**
M4		β_1^T	1.2	1.302	8.523	0.912	0.191	0.201
		β_2^T	- 1.2	- 1.296	7.996	0.874	0.110	0.105
Indep		β_1^C	- 1.4	- 1.804	28.842	0.322	0.168	0.166
		β_2^C	1.4	1.786	27.571	0.040	0.107	0.108
$\overline{\text{BIC}}$	1645.8	β_0	- 0.6	- 0.797	32.839	0.696	0.134	0.129
$\overline{\text{HQ}}$	1602.3	β_1	0.7	0.805	15.014	0.896	0.152	0.150
$\overline{\text{AIC}}$	1574.2	β_2	0.8	0.905	13.162	0.756	0.082	0.081
M4		β_1^T	1.2	1.281	6.712	0.928	0.183	0.188
		β_2^T	- 1.2	- 1.278	6.483	0.906	0.105	0.099
Under		β_1^C	- 1.4	- 1.631	16.523	0.690	0.147	0.145
		β_2^C	1.4	1.626	16.108	0.368	0.098	0.095
$\overline{\text{BIC}}$	1617.6	β_0	- 0.6	- 0.685	14.219	0.924	0.126	0.120
$\overline{\text{HQ}}$	1574.1	β_1	0.7	0.776	10.905	0.932	0.140	0.136
$\overline{\text{AIC}}$	1546.0	β_2	0.8	0.874	9.282	0.854	0.078	0.076
M4		β_1^T	1.2	1.166	- 2.854	0.922	0.159	0.166
		β_2^T	- 1.2	- 1.170	- 2.493	0.938	0.094	0.091
Correct		β_1^C	- 1.4	- 1.371	- 2.076	0.924	0.119	0.123
		β_2^C	1.4	1.374	- 1.852	0.948	0.086	0.084
$\overline{\text{BIC}}$	1599.9	β_0	- 0.6	- 0.585	- 2.436	0.936	0.109	0.110
$\overline{\text{HQ}}$	1556.4	β_1	0.7	0.685	- 2.075	0.948	0.113	0.119
$\overline{\text{AIC}}$	1528.2	β_2	0.8	0.775	- 3.152	0.918	0.069	0.073
M4		β_1^T	1.2	1.074	- 10.526	0.798	0.140	0.170
		β_2^T	- 1.2	- 1.081	- 9.904	0.666	0.087	0.096
Over		β_1^C	- 1.4	- 1.242	- 11.250	0.648	0.106	0.127
		β_2^C	1.4	1.248	- 10.882	0.514	0.081	0.087
$\overline{\text{BIC}}$	1619.7	β_0	- 0.6	- 0.571	- 4.889	0.878	0.100	0.118
$\overline{\text{HQ}}$	1576.2	β_1	0.7	0.627	- 10.420	0.794	0.096	0.128
$\overline{\text{AIC}}$	1548.1	β_2	0.8	0.710	- 11.260	0.624	0.063	0.081

To complement the evaluation of the estimates we present the AIC in Figure 1 for the four models and four scenarios. The M1 model in the *correct* scenario has the lowest median among all models and scenarios, as expected for the data generator model. Other models when fitted in the correct scenario present AIC similar to M1.

The *correct* scenario presents the lowest median among all scenarios, the median AIC increases as the dependency misspecification increases. In all models the *Under* scenario and the *Over* scenario present almost the same AIC median. The independent scenario presents a higher median than any other scenario for the four models. This shows that AIC and log-likelihood might be used to help determine the copula parameter (Figure 1).

Figure 1 - Akaike information criterion for the four models and four scenarios

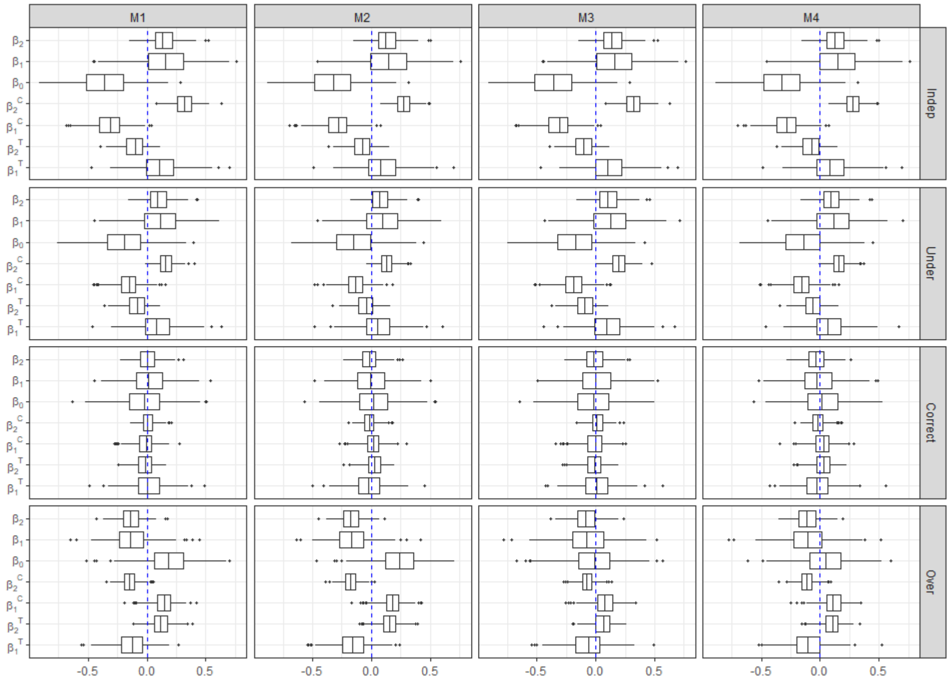


The relative bias is presented in Figure 2 for each parameter for all models scenarios. Figure 2 shows the bias reduction in the estimates as the dependency parameter is rightly set. The independent scenario has the greatest bias. The Figure also shows that even with the change of the copula function or the marginal model, the estimates are similar in terms of bias direction and interquartile range.

The independent case for β_2, β_1,β_2^C and β_1^T present median positive relative bias while β_0, β_1^C and β_2^T present median negative relative bias (Figure 2). These relative biases get closer to zero in the *Under* scenario and even more so in the *correct* scenario. In the *Over* scenario these biases get further away from zero, but in the other direction. This can be observed for the four models. Figure 2 also presents the variability of the estimates, β_0 and β_1 present the largest interquartile range, while β_2^C and β_1^T present the smallest.

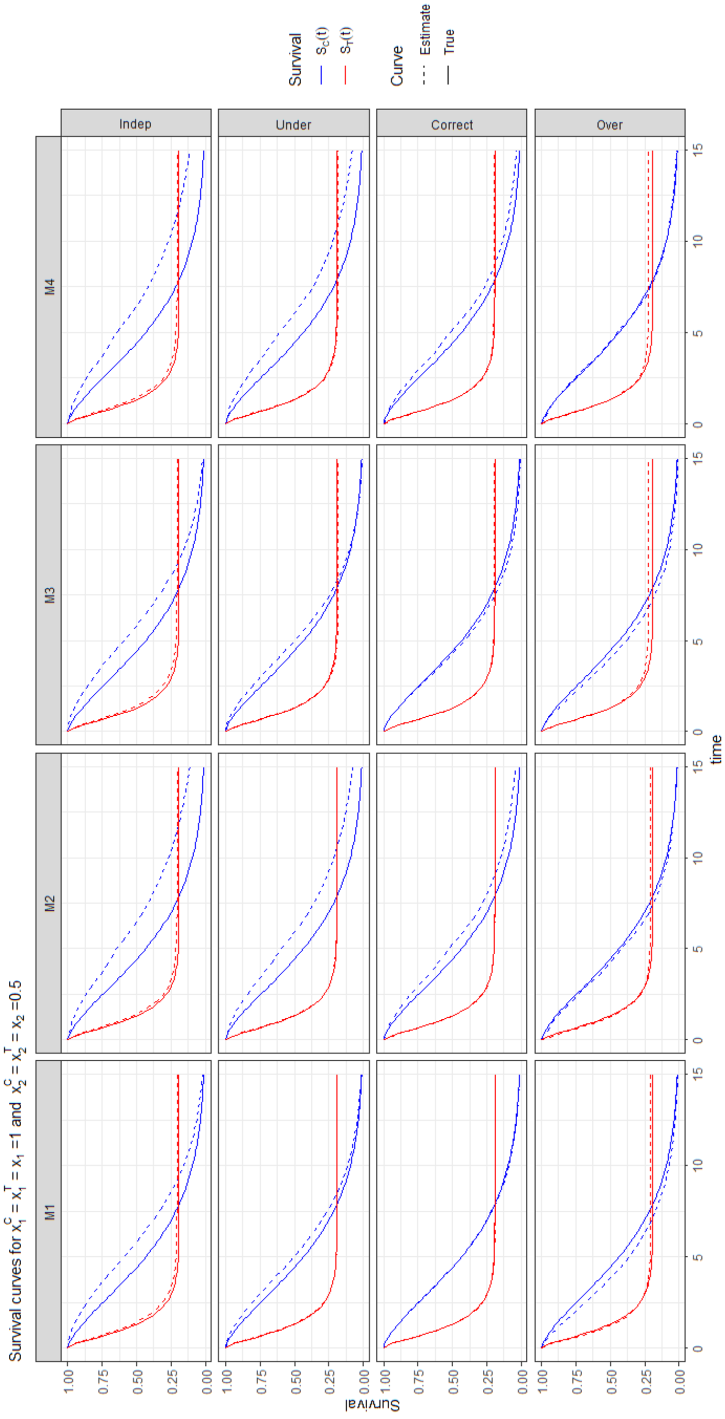
Figure 3 presents the survival curves of T and C for each model and scenario, to asses the impact of the estimates on the model's survival curve. The survival curves are obtained from the mean estimates of the parameters for $x_1^T = x_1^C = x_1 = 1$

Figure 2 - Parameters estimate's relative bias by model and scenario



and $x_2^T = x_2^C = x_2 = 0.5$. First, we note that the cure rate that can be seen in the curves. The red lines represent estimated survival function for T and it flattens out around 0.23 for all models, this means that cure probability is around 0.23 (Figure 3). Comparing all models and scenarios we see that the estimated survival curves overestimate the true curve when dependency is set bellow true value, and underestimates the true curve when dependency is set above the true dependency. This happens both for T and C but seems to be more drastic for C.

Figure 3 - Survival Curves by model and scenario



4 Analysis of prostate cancer data

Medical literature and review papers suggest that prostate cancer and cardiovascular disease might be correlated. According to Li et al. (2007) high cholesterol is a risk factor for both prostate cancer and cardiovascular disease. The prostate cancer treatment is often based on androgen deprivation therapies which have been associated with various cardiovascular diseases (Cardwell et al., 2020).

With this in mind we conduct the analysis of the SEER (National Institutes of Health surveillance epidemiology and end results) Prostate cancer data. The dataset, from which we randomly selected 25,000 cases are from patients diagnosed with prostate cancer in the year of 2000. From these 25,000 patients, 3,005(12.2%) passed away due the prostate cancer, 3,338(7.1%) due to heart diseases and the remaining 18,657(74.6%) are either alive ate the end of study or passed away due to other causes.

The variables selected to enter this study are the continuous age at diagnoses, that ranges form 40 to 90, with mean of 67.9 years and standard deviation of 9.40. We used the standardized age in the adjustment. We consider Race, dichotomized in White and Non white. Marital status dichotomized in yes and no, Surgery status with three classes: None for patients that did not underwent surgery; Resec which specifies patients that had tissue resection surgery; Destruc that specifies patients that had tissue destruction surgery. Lastly cancer stage, I if the information is not sufficient to assign a stage, II or III for an invasive neoplasm confined entirely to the prostate and IV if a neoplasm that has spread to other parts of the body.

Table 7 presents the prostate cancer dataset summary. **Count** denotes the number of patients that experienced each outcome, **(%)** denotes the percentage of patients that experienced the outcome and **Med** represents the median survival time of those patients. Is also presented the the total numbers and totals for all variable.

We assumed Alive or Other causes of death to be a right censoring, that is independent from the prostate cancer survival time. Heart disease we consider dependent censoring of the prostate cancer survival time. We will fit the four models to this dataset (M1, M2, M3 and M4). To choose which variables will model each outcome and the cure fraction we conducted marginal analysis. For the prostate cancer survival time all selected variables are significant and therefore will be used to model the prostate cancer. For the marginal fit of heart disease time, are significant age and marital status, we dichotomized surgery further into yes and no, which made this new variable significant. For the cure fraction we selected age and marital status.

So for M2 and M4 models we assumed grids with 5 intervals to adjust prostate cancer and heart disease. We selected 5 intervals as a more conservative view point, Li et al. (2016) uses 3 intervals and Loeb et al. (2011) uses 2, we understand that with 5 intervals we allow the model to be more flexible and therefore fit better the dataset. In order to fit the cure model with dependent censoring under copula approach we set Kendall's $\tau = 0.4$, based on Escarela and Carriere (2003) that

Table 7 - Summary and distribution of the variables included in study and comparison between Prostate Cancer. Heart disease and Alive/Other death causes

	Prostate Cancer			**Heart diseases**			**Alive/Other deaths**	
	Count	**(%)**	**Med**	**Count**	**(%)**	**Med**	**Count**	**(%)**
Total	3005	12.02	5	3338	13.352	7.1	18657	74.628
Age								
40-60	441	7.8	6.3	188	3.3	8.7	5039	88.9
61-68	656	9.4	6.5	554	7.9	8.5	5773	82.7
69-75	807	12	5.8	1098	16.4	7.8	4796	71.6
76-90	1101	19.5	3.4	1498	26.5	5.9	3049	54
Race								
White	2355	11.5	5.2	2768	13.5	7.3	15327	74.9
N White	650	14.3	4.7	570	12.5	6.3	3330	73.2
Marital Status								
Yes	1847	10.7	5.3	2142	12.4	7.4	13302	76.9
No	1158	15	4.5	1196	15.5	6.5	5355	69.5
Surgery								
None	2332	14.8	4.8	2584	16.4	7.1	10797	68.7
Ressec	508	6.1	6.4	572	6.8	7.8	7301	87.1
Destruc	165	18.2	4.3	182	20.1	5.3	559	61.7
Cancer Stage								
I	345	26.7	4.4	213	16.5	5.5	732	56.7
II or III	1951	8.6	6.8	3032	13.4	7.3	17715	78
IV	709	70.1	1.7	93	9.2	2.1	210	20.8

proved an identifiable model using Frank's copula that estimated $\nu = 4.28$ as the dependency among Prostate cancer and heart diseases, which relates to $\tau = 0.41$. So, we take the Clayton copula $\nu = 1.35$ and Plackett copula $\nu = 7.03$. We also do not provide adjustment diagnosis, in a more robust health data study residual analysis is recommended.

The Tables 8 ,9, 10 and 11 present the models regression coefficients estimates compared to the independent scenario estimates. In the tables **Est** stands for estimate, **SE** for standard error, **LCL** and **UCL** for lower and upper asymptotic Wald confidence interval limit, and **Diff est** is the dependent and independent model's estimates difference. We do not provide proof of asymptotic theory in this work as it is still lacking in the literature. asymptotic theory is well formulated for the cure models, but we could not find much on copula survival models, simulation study showed good performance of the Wald asymptotic confidence intervals and therefore we use it in our real data application.

Table 8 presents the regression coefficients estimates produced by M1 for $\nu = 1.35$ and compared to the independent scenario estimates. There is some differences in the estimates, Marital status, surgery (destructive) and stage(IV) present higher estimates in the independent model for the prostate cancer survival time. Surgery (destructive) is not significant in the dependent model. The signs of parameters

Table 8 - M1 model adjustment of Prostate cancer dataset

	M1 Model								
	Setting $\nu = 1.35$				Independent				**Diff est**
Parameter	**Est**	**SE**	**LCL**	**UCL**	**Est**	**SE**	**LCL**	**UCL**	
age	0.495	0.079	0.340	0.649	0.420	0.072	0.278	0.562	0.075
race white	-0.089	0.035	-0.157	-0.021	-0.171	0.050	-0.269	-0.074	0.082
marr yes	-0.713	0.178	-1.062	-0.365	-0.383	0.136	-0.650	-0.115	-0.331
surg resc	-0.379	0.041	-0.459	-0.300	-0.482	0.055	-0.589	-0.375	0.103
surg destr	0.013	0.066	-0.117	0.143	0.230	0.091	0.052	0.408	-0.217
stage II/III	-0.814	0.050	-0.911	-0.717	-1.267	0.066	-1.395	-1.138	0.453
stage IV	2.170	0.075	2.023	2.317	2.366	0.082	2.205	2.527	-0.196
age	0.941	0.018	0.906	0.975	1.095	0.023	1.051	1.139	-0.154
marr yes	-0.259	0.029	-0.317	-0.202	-0.221	0.036	-0.293	-0.150	-0.038
surg2 yes	-0.311	0.035	-0.380	-0.243	-0.266	0.043	-0.350	-0.181	-0.046
β_0	0.544	0.103	0.342	0.747	0.702	0.095	0.515	0.888	-0.157
age	0.160	0.074	0.014	0.306	0.078	0.066	-0.052	0.209	0.081
marr yes	0.419	0.166	0.093	0.745	0.107	0.125	-0.138	0.351	0.312

however are the same, both models find the same effect direction to each variable. For Heart diseases all parameters present lower values when considering dependency and all parameters are significant in both models. For the prostate cancer cure rate the independent model does not find significance for age and marital status, whereas the dependent model does. When comparing the standard errors in prostate cancer, only age and marital status present higher values for the dependent model. All parameters of heart diseases have lower standard error in the dependent model. All estimates from the independent model present lower standard error then the dependent model.

Table 9 presents the regression coefficients estimates produced by M3 for $\nu = 7.04$ and compared to the independent scenario estimates. M3 uses the Plackett copula to adjust dependency and assumes Weibull marginals. Analyzing difference in the estimates of the dependent model and the independent model we see that Age and stage II/III present higher estimates in the dependent model, for the prostate cancer survival time, all other parameters all smaller in the dependent model.As for Heart diseases all parameters present lower values when considering dependency (Table 9). For M3, Surgery (destructive) is not significant in the dependent model. In the independent model all parameters (from Prostate cancer and heart disease) are significantly different from zero. For the cure rate age and marital status are not significant under the independent model (Table 9). For the prostate cancer cure rate, the independent M3 model does not find significancy for age and marital status, as for the dependent M3 model the significancy is found. Only β_0 is lower in the dependent model, all other parameters present higher values (Table 9).

Table 10 presents the regression coefficients estimates produced by M2 for $\nu = 1.35$ and compared to the independent scenario estimates. Analyzing differences in the estimates of the dependent model and the independent model we see that

Table 9 - M3 model adjustment of Prostate cancer dataset

	M3 Model								
	Setting $\nu = 7.04$				Independent				**Diff est**
Parameter	**Est**	**SE**	**LCL**	**UCL**	**Est**	**SE**	**LCL**	**UCL**	
age	0.477	0.068	0.344	0.610	0.444	0.074	0.300	0.589	0.033
race white	-0.157	0.044	-0.244	-0.069	-0.151	0.050	-0.249	-0.053	-0.006
marr yes	-0.687	0.141	-0.962	-0.411	-0.376	0.137	-0.645	-0.107	-0.311
surg resc	-0.494	0.052	-0.596	-0.392	-0.472	0.055	-0.579	-0.365	-0.021
surg destr	0.101	0.081	-0.058	0.259	0.268	0.090	0.092	0.445	-0.168
stage II/III	-1.015	0.057	-1.128	-0.902	-1.246	0.066	-1.376	-1.117	0.231
stage IV	2.335	0.078	2.183	2.487	2.388	0.082	2.227	2.550	-0.053
age	1.071	0.021	1.030	1.112	1.095	0.023	1.051	1.139	-0.024
marr yes	-0.243	0.034	-0.309	-0.177	-0.233	0.036	-0.304	-0.161	-0.010
surg2 yes	-0.309	0.040	-0.387	-0.230	-0.250	0.043	-0.334	-0.166	-0.059
β_0	0.492	0.087	0.323	0.662	0.718	0.097	0.527	0.909	-0.226
age	0.153	0.063	0.030	0.276	0.057	0.068	-0.076	0.190	0.096
marr yes	0.367	0.129	0.115	0.620	0.111	0.125	-0.135	0.357	0.256

Age and stage II/III present higher estimates in the dependent model for the prostate cancer survival time, all other parameters are smaller in the dependent model. Surgery (destructive) is not significant in the dependent M2 model, in the independent model all parameters related to prostate cancer survival time and heart disease survival time are significantly different from zero (Table 10). The prostate cancer cure rate of the independent M2 model does not find significance for age and marital status, in the dependent model significance is found for all variables. β_0 estimate is smaller in the dependent model than the dependent model, all other parameters present higher values in the dependent model (Table 10).

Table 10 - M2 model adjustment of Prostate cancer dataset

	M2 Model								
	Setting $\nu = 1.35$				Independent				**Diff est**
Parameter	**Est**	**SE**	**LCL**	**UCL**	**Est**	**SE**	**LCL**	**UCL**	
age	0.495	0.045	0.407	0.582	0.486	0.049	0.389	0.583	0.009
race white	-0.080	0.059	-0.196	0.037	-0.302	0.086	-0.472	-0.133	0.222
marr yes	-0.811	0.134	-1.073	-0.549	-0.735	0.128	-0.986	-0.484	-0.076
surg resc	-0.438	0.071	-0.577	-0.299	-0.332	0.068	-0.465	-0.200	-0.106
surg destr	0.017	0.098	-0.174	0.209	0.313	0.099	0.119	0.506	-0.295
stage II/III	-0.803	0.075	-0.950	-0.656	-1.342	0.071	-1.482	-1.202	0.539
stage IV	2.022	0.080	1.866	2.179	1.998	0.070	1.860	2.135	0.024
age	0.824	0.023	0.780	0.868	0.975	0.025	0.927	1.023	-0.151
marr yes	-0.447	0.043	-0.530	-0.363	-0.383	0.040	-0.460	-0.305	-0.064
surg2 yes	-0.281	0.075	-0.428	-0.134	-0.235	0.076	-0.385	-0.085	-0.046
β_0	0.466	0.077	0.316	0.616	0.732	0.087	0.561	0.903	-0.266
age	0.129	0.056	0.020	0.238	0.011	0.056	-0.099	0.120	0.119
marr yes	0.549	0.119	0.317	0.782	0.102	0.123	-0.138	0.343	0.447

Table 11 presents the regression coefficients estimates produced by M4 for $\nu = 7.04$ and compared to the independent scenario estimates. Marital status, surgery (destructive) and stage IV present higher estimates in the independent model for the prostate cancer survival time, all other parameters are smaller in the dependent model. Looking at the confidence interval we see that Surgery (destructive) is not significant in the dependent M4 model, in the independent model all parameters are significantly different from zero (Table 11). For Heart diseases all parameters present lower values when considering dependent M4 compared to the independent model, all parameters are significant. Under the independent M4 model for prostate cancer's cure rate, age and marital status are not significant, as for the dependent model significance is found for all variables related to the cure rate. Moreover, only β_0 is smaller in the dependent model, all other parameters present higher values in this model (Table 11).

Table 11 - M4 model adjustment of Prostate cancer dataset

	M4 Model								
	Setting $\nu = 7.04$				Independent				**Diff est**
Parameter	**Est**	**SE**	**LCL**	**UCL**	**Est**	**SE**	**LCL**	**UCL**	
age	0.481	0.041	0.400	0.562	0.469	0.047	0.377	0.560	0.013
race white	-0.163	0.066	-0.292	-0.034	-0.263	0.085	-0.429	-0.097	0.101
marr yes	-0.821	0.125	-1.066	-0.575	-0.714	0.118	-0.945	-0.483	-0.107
surg resc	-0.293	0.061	-0.413	-0.173	-0.315	0.067	-0.445	-0.185	0.022
surg destr	0.125	0.104	-0.079	0.329	0.277	0.101	0.080	0.474	-0.152
stage II/III	-0.963	0.064	-1.088	-0.837	-1.344	0.071	-1.484	-1.205	0.382
stage IV	2.002	0.071	1.864	2.140	2.054	0.067	1.923	2.184	-0.052
age	0.965	0.024	0.918	1.012	0.989	0.025	0.941	1.037	-0.024
marr yes	-0.480	0.038	-0.555	-0.405	-0.390	0.039	-0.466	-0.313	-0.090
surg2 yes	-0.298	0.067	-0.430	-0.167	-0.243	0.076	-0.392	-0.094	-0.055
β_0	0.551	0.068	0.417	0.685	0.729	0.087	0.559	0.900	-0.178
age	0.184	0.047	0.092	0.275	0.028	0.054	-0.079	0.134	0.156
marr yes	0.708	0.102	0.507	0.909	0.109	0.116	-0.118	0.336	0.599

We see that all four models reach similar conclusions when comparing the dependent scenarios. The regression coefficients estimates are similar between models, this shows that the Weibull and piecewise exponential model were able to fit the data and agree on the regression coefficients and their significance. Both copula functions produced similar results, reinforcing that the selection of the copula parameter is more important than the copula function.

Comparing the independent scenarios the same behavior is seen. All models found similar estimates for the parameters and agree on their significance. Copula function and marginal models did not have great influences on the estimates, as they are similar throughout the models.

There might be some model specification problem that arise from the fact that we do not have much information on general health status of the prostate cancer patient, for example diabetes, cholesterol and blood pressure, known to be important risk factor.

5 Discussion and conclusion

In this paper we constructed cure rate models for dependent censoring under copula approach. We used the non-mixture method to build models that allows a cure rate. To account the dependency among the time to event of interest and time to dependent censoring we used copula functions. Our study focused on two copula functions: Clayton and Plackett, to adjust the dependency between lifetime and dependent censoring. We focused on two marginal time distributions, Weibull and piecewise exponential distributions.

With the proposed models we proceeded a simulation study. The simulation study showed the models are reliable in analyzing survival times with cure rate, as well as, with dependency between lifetime and censoring time. The simulation study reinforced the importance of the dependency modeling, and additionally, showed the estimation bias that raises from the independence assumption. The simulations also showed that the cure rate model for dependent censoring under copula approach estimates, in general, present small relative bias, small standard errors and adequate coverage probabilities when the dependency is properly set.

The piecewise exponential distribution was able to adjust the generated datasets, which were built from the Weibull distribution. Furthermore, it find regression coefficient estimates with small bias and good coverage probability. The copulas function choice seems to be less important than the copulas parameter.

After the simulations, we conducted the adjustment to the Prostate cancer dataset. We assumed a fixed copula parameter, Kendall's correlation coefficient τ was set to 0.4. We also fit the models assuming independency to compare the results with the dependent assumptions.

The dependent models found similar estimates for the parameters and agreed on the significances. The same behavior can be seen in the independent scenario, all the models found similar estimates and agreed on their significances. Comparing the adjustments under the the independent scenario and the adjustments under the dependent scenario the parameters estimates differed. Models in the dependent scenario found more parameters to be significant, specially those related to the cure rate. Models under the dependent scenario present most parameters with smaller estimates then those found under independent scenario.

The proposed models are a general formulation, that can easily be extended to accommodate other baseline parametric distributions or non-parametric distributions, as well as, other copula functions. In this paper we do not estimate copula parameter ν that adjust the dependency between time to event of interest and dependent censoring time, instead we use a fixed value for the prostate cancer dataset and use dependency scenarios in the simulations. The estimation of the copula parameters remains as a future goal.

References

Almetwally, E. M., Almongy, H. M., and El sayed Mubarak, A. (2018). Bayesian and maximum likelihood estimation for the weibull generalized exponential distribution parameters using progressive censoring schemes. *Pakistan Journal of Statistics and Operation Research*, pages 853–868.

Boag, J. W. (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.

Cancho, V. G., Rodrigues, J., and de Castro, M. (2011). A flexible model for survival data with a cure rate: a bayesian approach. *Journal of Applied Statistics*, 38(1):57–70.

Cardwell, C. R., Sullivan, J. M., Jain, S., Harbinson, M. T., Cook, M. B., Hicks, B. M., and McMenamin, Ú. C. (2020). The risk of cardiovascular disease in prostate cancer patients receiving androgen deprivation therapies. *Epidemiology*, 31(3):432–440.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):235–251.

Clayton, D. G. (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.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Emura, T. and Chen, Y.-H. (2018). *Analysis of survival data with dependent censoring: Copula-based approaches.* Springer.

Emura, T. and Michimae, H. (2017). A copula-based inference to piecewise exponential models under dependent censoring, with application to time to metamorphosis of salamander larvae. *Environmental and ecological statistics*, 24(1):151–173.

Escarela, G. and Carriere, J. F. (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4):333–349.

Han, B., Van Keilegom, I., and Wang, X. (2021). Semiparametric estimation of the nonmixture cure model with auxiliary survival information. *Biometrics.*

Hsu, T.-M., Emura, T., and Fan, T.-H. (2016). Reliability inference for a copula-based series system life test under multiple type-i censoring. *IEEE Transactions on Reliability*, 65(2):1069–1080.

Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics*, 58(3):510–520.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics*, 64(4):1090–1099.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2016). *Handbook of survival analysis*. CRC Press.

Lambert, P. and Bremhorst, V. (2019). Estimation and identification issues in the promotion time cure model when the same covariates influence long-and short-term survival. *Biometrical Journal*, 61(2):275–289.

Lambert, P. and Bremhorst, V. (2020). Inclusion of time-varying covariates in cure survival models with an application in fertility studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1):333–354.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

Li, D., Hu, X. J., McBride, M. L., and Spinelli, J. J. (2019). Multiple event times in the presence of informative censoring: modeling and analysis by copulas. *Lifetime data analysis*, pages 1–30.

Li, Y., Panagiotou, O. A., Black, A., Liao, D., and Wacholder, S. (2016). Multivariate piecewise exponential survival modeling. *Biometrics*, 72(2):546–553.

Li, Y., Tiwari, R. C., and Guha, S. (2007). Mixture cure survival models with dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):285–306.

Liu, Y., Hu, T., and Sun, J. (2017). Regression analysis of current status data in the presence of a cured subgroup and dependent censoring. *Lifetime data analysis*, 23(4):626–650.

Loeb, S., Vonesh, E. F., Metter, E. J., Carter, H. B., Gann, P. H., and Catalona, W. J. (2011). What is the true number needed to screen and treat to save a life with prostate-specific antigen testing? *Journal of Clinical Oncology*, 29(4):464.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Palaro, H. P. and Hotta, L. K. (2006). Using conditional copula to estimate value at risk. *Journal of Data Science*, 4:93–115.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, 60(310):516–522.

Rondeau, V., Schaffner, E., Corbiere, F., Gonzalez, J. R., and Mathoulin-Pélissier, S. (2013). Cure frailty models for survival data: Application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Statistical methods in medical research*, 22(3):243–260.

Rowley, M., Garmo, H., Van Hemelrijck, M., Wulaningsih, W., Grundmark, B., Zethelius, B., Hammar, N., Walldius, G., Inoue, M., Holmberg, L., et al. (2017). A latent class model for competing risks. *Statistics in medicine*, 36(13):2100–2119.

Salvadori, G., De Michele, C., Kottegoda, N. T., and Rosso, R. (2007). *Extremes in nature: an approach using copulas*, volume 56. Springer Science & Business Media.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656.

Siannis, F. (2004). Applications of a parametric model for informative censoring. *Biometrics*, 60(3):704–714.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

Wang, S., Xu, D., Wang, C., and Sun, J. (2021). Semiparametric analysis of case k interval-censored failure time data in the presence of a cured subgroup and informative censoring. *Journal of Statistical Computation and Simulation*, pages 1–16.

Wey, A., Salkowski, N., Kremers, W., Ahn, Y. S., and Snyder, J. (2020). Piecewise exponential models with time-varying effects: Estimating mortality after listing for solid organ transplant. *Stat*, 9(1):e264.

William, J. S. and Lagakos, S. W. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. *Biometrika*, 64(2):215–224.

Yakovlev, A. Y., Asselain, B., Bardou, V., Fourquet, A., Hoang, T., Rochefediere, A., and Tsodikov, A. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et analyse de donnees spatio-temporelles*, 12:66–82.

Yakovlev, A. Y., Cantor, A. B., and Shuster, J. J. (1994). Parametric versus nonparametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, 13(9):983–986.

Yakovlev, A. Y. and Tsodikov, A. (1996). Stochastic models of tumor latency and their biostatistical applications.

Zhang, Z., Sun, L., Sun, J., and Finkelstein, D. M. (2007). Regression analysis of failure time data with informative interval censoring. *Statistics in medicine*, 26(12):2533–2546.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138.

Received on 0x.0x.20xx.

Approved after revised on 0x.0x.20xx.