Use of Bivariate Lifetime Distributions Assuming Continuous or Discrete Data Applied to Patients with Breast Cancer

Tatiana Reis Icuma[1], Isabela Panzeri Carlotti Buzatto[2], Daniel Guimarães Tiezzi[2],

Jorge Alberto Achcar[1*], Nasser Davarzani[3]

[1] Department of Social Medicine, Medical School, University of São Paulo,
Ribeirão Preto, SP, Brazil
[2]Department of Obstetrics and Gynecology, University of São Paulo,
Ribeirão Preto, SP, Brazil
[3]Department of Knowledge Engineering, Maastricht, the Netherlands

Abstract: Breast cancer is the second most common type of cancer in the world (World Cancer Report, 2014 a, b). The evolution of breast cancer treatment usually allows a longer life of patients as well in many cases a relapse of the disease. Usually medical researchers are interested to analyze data denoting the time until the occurrence of an event of interest such as the time of death by cancer in presence of right censored data and some covariates. In some situations, we could have two lifetimes associated to the same patient, as for example, the time free of the disease until recurrence and the total lifetime of the patient. In this case, it is important to assume a bivariate lifetime distribution which describes the possible dependence between the two observations. We consider as an application, different parametric bivariate lifetime distributions to analyze a breast cancer data set considering continuous or discrete data. Inferences of interest are obtained under a statistical Bayesian approach. We get the posterior summaries of interest using existing MCMC (Markov Chain Monte Carlo) methods. The main goal of the study, is to compare the bivariate continuous and discrete distributions that better describes the breast cancer lifetimes.

Key words: Bayesian analysis, Bivariate lifetime distributions, Breast cancer data, Continuous and discrete lifetime data, Right censored data and covariates.

1. Introduction and motivation

It is estimated that every year about one and a half million new cases of breast cancer arise worldwide. It is the most common form of the disease in women, comprising 25% of all cancers (World Cancer Report, 2014 a, b). In the United States of America, it is the second largest cause of cancer deaths (DeSantis et al., 2014), and it is expected that every one in eight women will develop the disease in their lifetime (DeSantis et al., 2014). Breast cancer is a cancer that develops from breast tissue (Saunders and Jassal, 2009). Breast cancer signs can include a lump in the breast, a change in shape of the breast, dimpling of the skin, fluid coming from the nipple, or a red scaly patch on the skin. The medical literature contains a large number of papers related to breast cancer, risk factors and possible therapies (see for example, Lacroix, 2006; Reeder and Vogel, 2008; Yager and Davidson, 2006; Kahlenborn et al, 2006; Gaffield et al , 2009; Yang and Jacobsen, 2008; Brody et al, 2007; Ferro, 2012; Hendrix, 2010; Colditz et al, 2012, Nelson et al 2012; Sotiriou and Pusztai, 2009; Cuzick et al, 2013; Wu et al, 2008; Kelsey, 1993; McPherson et al, 2000; Tiezzi, 2009).

In oncology studies, often the researcher's interest is related to a variable representing the time until the occurrence of an event such as the time until relapse of disease or time to death. The statistical survival analysis methodology is therefore commonly used in the analysis of survival data from cancer clinical trials, with emphasis to the Cox proportional hazards model (Cox, 1972); to parametric regression models based on existing lifetimes distributions or to nonparametric methods for comparisons between survival curves, such as the Wilcoxon and log-rank tests (see, for example, Lee and Wenyuwang, 2003). The use of the Cox proportional hazards regression model (or the equivalent in the case of two treatment groups, the log-rank test) is well established as common practice in the statistical analysis of medical data. The great popularity of this model is motivated by being free of a parametric distribution assumption for the survival times.

In some situations we could have more than one response of interest associated to the same patient, which require parametrical models based on some bivariate distribution in place of usual medical approach assuming independent survival times.

In this paper, we consider as an application, a bivariate lifetime data set related to a study accomplished at the Clinics Hospital, Medical School of the University of São Paulo, Ribeirão Preto, Brazil related to 54 female patients with locally advanced breast cancer (stage II and III) with overexpressing HER-2 (HER-2 positive), submitted to neoadjuvant chemotherapy from 2008 to 2012 and where the researchers have two variables of interest associated to each patient: the time until relapse of the disease and the total time to death (data in Table 1A, Appendix A).

The main goal of this study is to verify whether there are significant differences in the patients' survival times that received at least four cycles of a drug Herceptin before surgery or less than four cycles. Trastuzumab (brand name of the drug Herceptin®) is a monoclonal antibody derived from recombinant DNA linked with high affinity Growth Factor Receptor Human Epidermal 2 (HER2). According to the European body "European Medicines Agency" Herceptin is used to treat various types of cancer: breast cancer in early stage (when the cancer has spread within the breast or to the glands under the arm but has not spread to other parts of the body) after surgery, chemotherapy or radiotherapy. It can also be used in a preliminary stage of treatment in combination with chemotherapy. In the case of locally advanced tumors (including inflammatory) or more than two centimeters wide, Herceptin is used before surgery in combination with chemotherapy, and again after surgery alone; cancer metastatic breast cancer (cancer that has spread to other parts of the body). It is used as monotherapy in patients who have not responded to previous treatments. The Herceptin is also used in combination with other cancer drugs. The Herceptin can also be used for gastric cancer (stomach cancer) breast cancer in combination with other medicines. The Herceptin can only be used after having been shown that cancer presents "overexpress" HER2.

As secondary goals, we have,

- The medical researchers want to check if different factors (age, stage, type of surgery, pathologic complete response, positivity to estrogen receptor, positivity to progesterone receptor or relapse) have significant effects on the survival times of the patients (DFS: disease-free survival and TS: total survival, both in months).
- To verify if different bivariate lifetime distributions could imply in different and better inferences of interest.
- To present a discussion on the choice of the best model.

For a statistical analysis of bivariate lifetimes in presence of covariates and censored data, the literature presents different parametrical models assuming continuous or discrete lifetime data. In this way, for the analysis of the breast cancer data of Table 1A, we could assume different bivariate lifetime models for the responses DFS and TS denoted as lifetimes T_1 and T_2 associated to each patient. Some popular models for bivariate lifetimes assume bivariate exponential distributions considering continuous data. Among these bivariate exponential distributions, one model has being extensively used by reliability engineers and medical researchers: the exponential bivariate model introduced by Block and Basu (1974). Many other continuous models also are introduced in the literature (see for example, Freund, 1988; Coelho-Barros et al., 2016; Louzada-Neto et al., 2012). Other possibility: the use of bivariate lifetime models for discrete data (see for example, Arnold, 1975; Basu and Dhar, 1995; Davarzani et al., 2015).

In each application usually the researchers should decide by one of the existing bivariate lifetime models in the analysis of the data, a task usually not simple.

The paper is organized as follows: in section 2, we present the dataset of the study and the bivariate lifetime distributions used in the data analysis; in section 3, we present the results of a Bayesian analysis for the cancer survival times of Table 1A assuming the three models introduced in section 2 in presence or not of covariates; finally in section 4, we introduce some conclusions and discussion of the obtained results.

2. Material and methods

The dataset is related to 54 female patients with stages II and III breast cancer with HER-2 overexpression (HER-2 positive) in the period ranging from 2008 to 2012, undergoing neoadjuvant chemotherapy associated with the drug Herceptin® . Each patient was followed up from the date of the first visit at the hospital until the end of the study.

Around 25% of all breast cancer overexpress the human epidermal growth factor receptor 2 (HER2) and is considered a more aggressive disease (Vu et al., 2012). The patients in this dataset received the drug trastuzumab (Herceptin®), which is a humanized monoclonal antibody that binds to HER-2. Some patients received at least 4 cycles of the drug before surgery, while others received less than four cycles pre-operatively. All patients received the drug after surgery for one year except those that developed severe toxicity. Data from this study shows two responses of interest associated with each patient: the disease-free survival time (DFS-the patient may relapse or not) and the overall survival time in months (TS-death or survival until the last follow-up). The columns "Relapse" and "Death" of Table 1A in Appendix A at the end of the paper contain the information of censorship associated respectively with the DFS (Disease free times) and the TS (Overall survival times).

In Table 1A, we have the following definitions for each column:
- Ident: Identification of the patient.
- Age: Age of the patient (in years).
- Hercep: Use of the drug Herceptin® (1: ≥ 4 cycles; 2: < 4 cycles).
- Stage: stage of the disease (2 or 3).
- Surgical: Type of surgery performed on Patient (1: radical; conservative: 0).
- pCR: pathologic Complete Response (1: Yes; 0: No).
- Estr: Estrogen Receptor (1: positive; 0: negative).
- Proge: Progesterone Receptor (1: positive; 0: negative).
- Relapse: Relapse (1: Yes; 0: No).
- DFS: Disease-free survival (in months).
- Death: Death of the patient, for breast cancer (1: Yes; 0: No).
- TS: Overall or total survival (in months)

To analyze the breast cancer data given in Table 1A, we assume three bivariate lifetimes distributions introduced in the literature: the Block and Basu distribution, the Arnold distribution and the Basu-Dhar distribution.

2.1. Dependent lifetimes assuming a Block and Basu bivariate exponential distribution

The bivariate exponential distribution of Block and Basu (1974) with parameters λ_1, λ_2 and λ_3 for the lifetimes T_1 and T_2 has a joint density function given by:

$$f(t_1,t_2) = \begin{cases} f_1(t_1,t_2) = \frac{\lambda\lambda_1\lambda_{23}}{\lambda_{12}} exp\{-\lambda_1 t_1 - \lambda_{23} t_2\} & ,if\ t_1 < t_2 \\ f_2(t_1,t_2) = \frac{\lambda\lambda_2\lambda_{13}}{\lambda_{12}} exp\{-\lambda_{13} t_1 - \lambda_2 t_2\} & ,if\ t_1 \geq t_2 \end{cases} \tag{1}$$

where, $\lambda_{12} = \lambda_1 + \lambda_2$, $\lambda_{13} = \lambda_1 + \lambda_3$, $\lambda_{23} = \lambda_2 + \lambda_3$ and $\lambda = \lambda_1 + \lambda_2 + \lambda_3$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ and $\lambda_3 \geq 0$.

The means and variances for T_1 and T_2 are given, respectively, by

$$\mu_1 = E(T_1) = \frac{1}{\lambda_{13}} + \frac{\lambda_2\lambda_3}{\lambda\lambda_2\lambda_{13}} \text{ and } \mu_2 = E(T_2) = \frac{1}{\lambda_{23}} + \frac{\lambda_1\lambda_3}{\lambda\lambda_{12}\lambda_{23}} \tag{2}$$

$$\sigma_1^2 = Var(T_1) = \frac{1}{\lambda_{13}^2} + \frac{\lambda_2\lambda_3(2\lambda_1\lambda+\lambda_2\lambda_3)}{\lambda^2\lambda_{12}^2\lambda_{13}^2} \text{ and } \sigma_2^2 = Var(T_2) = \frac{1}{\lambda_{23}^2} + \frac{\lambda_1\lambda_3(2\lambda_2\lambda+\lambda_1\lambda_3)}{\lambda^2\lambda_{12}^2\lambda_{23}^2}$$

The correlation coefficient for T_1 and T_2 is given by:

$$\rho_{12} = \frac{\lambda_3[(\lambda_1^2+\lambda_2^2)\lambda+\lambda_1\lambda_2\lambda_3]}{\phi_1\phi_2} \tag{3}$$

where,

$\phi_1 = [\lambda_{12}^2\lambda_{13}^2 + \lambda_2(\lambda_2 + 2\lambda_1)\lambda^2]^{1/2}$ and $\phi_2 = [\lambda_{12}^2\lambda_{23}^2 + \lambda_1(\lambda_1 + 2\lambda_2)\lambda^2]^{1/2}$

The covariance between T_1 and T_2 is given by:

$$\mathrm{Cov}(\mathrm{T}_1, \mathrm{T}_2) = \frac{(\lambda_1^2+\lambda_2^2)\lambda_3\lambda+\lambda_1\lambda_2\lambda_3^2}{\lambda^2\lambda_{12}\lambda_{13}\lambda_{23}} \tag{4}$$

Inferences for the Block and Basu distribution under a Bayesian approach are introduced by different authors (see for example, Achcar and Leandro, 1998; or Achcar and Santos, 2011).

Assuming T_1 or T_2 as censored data where the censorship is independent of the lifetimes, we sub-divide the n observations into four classes:

C_1: both $\mathrm{t}_{1\mathrm{i}}$ and $\mathrm{t}_{2\mathrm{i}}$ are observed lifetimes;
C_2: $\mathrm{t}_{1\mathrm{i}}$ is an observed lifetime and $\mathrm{t}_{2\mathrm{i}}$ is a censored time (that is, we only know that $\mathrm{T}_{2\mathrm{i}} \geq \mathrm{t}_{2\mathrm{i}}$);
C_3: $\mathrm{t}_{1\mathrm{i}}$ is a censored time and $\mathrm{t}_{2\mathrm{i}}$ is an observed lifetime;
C_4: both $\mathrm{t}_{1\mathrm{i}}$ and $\mathrm{t}_{2\mathrm{i}}$ are censored times, (5)

where i = 1,…,n.

The likelihood function for a continuous model (see for example, Lawless, 1982, page 479) is given by:

$$L = \prod_{i\in C_1} f(t_{1i}, t_{2i}) \prod_{i\in C_2}\left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}}\right)\prod_{i\in C_3}\left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}}\right)\prod_{i\in 4} S(t_1, t_2) \tag{6}$$

where,

- $-\frac{\partial S(\mathrm{t}_{1\mathrm{i}},\mathrm{t}_{2\mathrm{i}})}{\partial \mathrm{t}_{1\mathrm{i}}} = \begin{cases} \mathrm{S}'_{1\mathrm{t}_1}(\mathrm{t}_{1\mathrm{i}}, \mathrm{t}_{2\mathrm{i}}) & ,\text{if } \mathrm{t}_{1\mathrm{i}} < \mathrm{t}_{2\mathrm{i}} \\ \mathrm{S}'_{2\mathrm{t}_1}(\mathrm{t}_{1\mathrm{i}}, \mathrm{t}_{2\mathrm{i}}) & ,\text{if } \mathrm{t}_{1\mathrm{i}} \geq \mathrm{t}_{2\mathrm{i}} \end{cases}$

$S'_{1t_i}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_1}{\lambda_{12}}\exp\{-\lambda_1 t_1 - \lambda_{23} t_{2i}\}$ and

$$S'_{2t_i}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_{13}}{\lambda_{12}}\exp\{-\lambda_{13} t_{1i} - \lambda_2 t_{2i}\} - \frac{\lambda\lambda_3}{\lambda_{12}}\exp\{-\lambda t_{1i}\}$$

- $-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}} = \begin{cases} S'_{1t_2}(t_{1i}, t_{2i}) & , if\ t_{1i} < t_{2i} \\ S'_{2t_2}(t_{1i}, t_{2i}) & , if\ t_{1i} \geq t_{2i} \end{cases}$

$S'_{1t_2}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_{23}}{\lambda_{12}}\exp\{-\lambda_1 t_{1i} - \lambda_{23} t_{2i}\} - \frac{\lambda\lambda_3}{\lambda_{12}}\exp\{-\lambda t_{2i}\}$ and

$$S'_{2t_2}(t_{1i}, t_{2i}) = \frac{\lambda\lambda_2}{\lambda_{12}}\exp\{-\lambda_{13} t_{1i} - \lambda_2 t_{2i}\}$$

- $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2) = \begin{cases} S_1(t_1, t_2) & , if\ t_1 < t_2 \\ S_2(t_1, t_2) & , if\ t_1 \geq t_2 \end{cases}$

$$S_1(t_1, t_2) = \frac{\lambda}{\lambda_{12}} \exp(-\lambda_1 t_1 - \lambda_{23} t_2) - \frac{\lambda_3}{\lambda_{12}} \exp(-\lambda t_2)$$

$$S_2(t_1, t_2) = \frac{\lambda}{\lambda_{12}} \exp(-\lambda_{13} t_1 - \lambda_2 t_2) - \frac{\lambda_3}{\lambda_{12}} \exp(-\lambda t_1)\ . \qquad (7)$$

For a Bayesian analysis of the Block and Basu distribution in presence of censored observations, we assume independent gamma prior distributions for the parameters λ_k, namely:

$$\lambda_k \sim \text{Gamma}\ (a_k, b_k) \qquad (8)$$

for k = 1,2 and 3; a_k and b_k are known hyperparameters; Gamma (a_k, b_k) denotes a gamma distribution with mean $a_k\ /\ b_k$ and variance $a_k\ /\ b_k^2$.

In the presence of a covariate vector **x**, let us consider the model:

$$\lambda_{1i} = \alpha_1 \exp\{\boldsymbol{\beta}_1' \mathbf{x}_i\}$$

$$\lambda_{2i} = \alpha_2 \exp\{\boldsymbol{\beta}_2' \mathbf{x}_i\} \qquad (9)$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jp})'$; $j = 1, 2$ is the vector of regression parameters and $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{pi})$, i =1,2,...,n.

In this case, we assume the following prior distributions for the parameters $\alpha_1, \alpha_2, \beta_{1l}, \beta_{2l}$ and λ_3:

$$\alpha_k \sim \text{Gamma}\ (c_k, d_k);\ \ \beta_{kl} \sim N(0, \sigma_{kl}^2)\ \text{ and }\ \lambda_3 \sim \text{Gamma}(e, f) \qquad (10)$$

for k=1,2; l=1,2,...,p; c_k, d_k, e, f and σ_{kl}^2 are known hyperparameters and $N(0, \sigma_{kl}^2)$ denotes a normal distribution with mean equals to zero and variance equals to σ_{kl}^2. We further assume prior independence among all parameters.

Under the Bayesian framework, we use Markov Chain Monte Carlo (MCMC) methods (see for example, Casella and George, 1992; Chib and Greenberg, 1995; Gelfand and Smith, 1990) and the OpenBugs software (Spiegelhalter, et al, 2003) to simulate samples of the joint posterior distribution of interest. Using OpenBugs we do not need to use all the conditional posterior distributions required for the Gibbs sampler algorithm (not included in this article) and we only need the likelihood function and prior distributions for the model parameters. From the simulated Gibbs samples, we get Monte Carlo estimates for the posterior summaries of interest.

2.2. Dependent lifetimes assuming an Arnold bivariate geometric distribution

An alternative for the use of a continuous distribution for the bivariate survival times is to assume the lifetimes T_1 and T_2 as discrete random variables that can take values on any positive integer approaching the decimal part of the survival time to the nearest integer.

In this way, the literature introduces different discrete bivariate distributions that could be used to analyze the bivariate lifetime data of Table 1A. A special multivariate discrete distribution was introduced by Arnold (1975) motivated from a Marshall-Olkin multivariate exponential distribution. In another direction, Nair and Nair (1988) studied the characteristics of some geometric bivariate exponential distributions. The bivariate geometric distribution proposed by Arnold (1975) has the probability mass function given by:

$$P(T_1 = t_1, T_2 = t_2) = \begin{cases} P_1(t_1, t_2) = \theta_1\theta_2(1-\theta_1-\theta_2)^{t_1-1}(1-\theta_2)^{t_2-t_1-1}, & t_1 < t_2 \\ 0 & , t_1 = t_2 \\ P_2(t_1, t_2) = \theta_1\theta_2(1-\theta_1-\theta_2)^{t_2-1}(1-\theta_1)^{t_1-t_2-1}, & t_1 > t_2 \end{cases} \qquad (11)$$

where the marginal mass probability functions for T_1 and T_2 are standard geometric distributions starting at one, given, respectively by,

$$p(t_1) = (1-\theta_1)^{t_1-1}\theta_1 \;, t_1 = 1,2,3,\ldots \text{ and } p(t_2) = (1-\theta_2)^{t_2-1}\theta_2 \;, t_2 = 1,2,3,\ldots$$

The means, variances, covariance and correlation are given, respectively, by,

$$\mu_1 = E(T_1) = \frac{1}{\theta_1} \text{ and } \mu_2 = E(T_2) = \frac{1}{\theta_2}$$

$$\sigma_1^2 = Var(T_1) = \frac{1-\theta_1}{\theta_1^2} \text{ and } \sigma_2^2 = Var(T_2) = \frac{1-\theta_2}{\theta_2^2}$$

$$Cov(T_1, T_2) = \frac{-1}{1-r} \text{ and } \rho_{12} = Corr(T_1, T_2) = -\frac{\theta_1\theta_2}{(1-r)[(1-\theta_1)(1-\theta_2)]^{0.5}} \qquad (12)$$

where $= 1-\theta_1-\theta_2$, $0 < \theta_1 < 1$ and $0 < \theta_2 < 1$.

Assume Y_1 and Y_2 as the censoring indicators of T_1 and T_2 which are independent of the lifetimes. As considered for the continuous case (see (5)), we also sub-divide the n observations into four classes:

C_1: $T_{1i} < Y_{1i}$ and $T_{2i} < Y_{2i}$, that is, t_{1i} and t_{2i} are complete lifetimes;
C_2: $T_{1i} < Y_{1i}$ and $Y_{2i} < T_{2i}$, thus we observe t_{1i} and t_{2i} is censored;
C_3: $Y_{1i} < T_{1i}$ and $T_{2i} < Y_{2i}$, thus we observe t_{2i} and t_{1i} is censored;
C_4: $Y_{1i} < T_{1i}$ and $Y_{2i} < T_{2i}$, thus t_{1i} and t_{2i} are both censored data. (13)

Given the above definitions, the likelihood function for θ_1 and θ_2 assuming a Arnold bivariate geometric distribution with probability mass function (11) with right censored data is given by:

$$L(\theta_1,\theta_2)=\prod_{i\in C_1}P(t_{1i},t_{2i})\prod_{i\in C_2}\left(\sum_{t_{2i}=y_{2i+1}}^{\infty}P(t_{1i},t_{2i})\right)\prod_{i\in C_3}\left(\sum_{t_{1i}=y_{1i+1}}^{\infty}P(t_{1i},t_{2i})\right)$$
$$\prod_{i\in C_4}\left(\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty}P(t_{1i},t_{2i})\right) \qquad (14)$$

where,

- $\sum_{t_{2i}=y_{2i+1}}^{\infty}P(t_{1i},t_{2i})=\begin{cases}\sum_{t_{2i}=y_{2i+1}}^{\infty}P_1(t_{1i},t_{2i}) \ , if\ t_{1i}<t_{2i}\\ \sum_{t_{2i}=y_{2i+1}}^{\infty}P_2(t_{1i},t_{2i}) \ , if\ t_{1i}\geq t_{2i}\end{cases}$

$$\sum_{t_{2i}=y_{2i+1}}^{\infty}P_1(t_{1i},t_{2i})=\theta_1(1-\theta_1-\theta_2)^{t_{1i}-1}(1-\theta_2)^{y_{2i}-t_{1i}-1}$$

$$\sum_{t_{2i}=y_{2i+1}}^{\infty}P_2(t_{1i},t_{2i}) \ =\theta_1(1-\theta_1-\theta_2)^{y_{2i}-1}(1-\theta_1)^{t_{1i}-y_{2i}-1}$$

- $\sum_{t_{1i}=y_{1i+1}}^{\infty}P(t_{1i},t_{2i})=\begin{cases}\sum_{t_{1i}=y_{1i+1}}^{\infty}P_1(t_{1i},t_{2i}) \ , if\, t_{1i}<t_{2i}\\ \sum_{t_{1i}=y_{1i+1}}^{\infty}P_2(t_{1i},t_{2i}) \ , if\, t_{1i}\geq t_{2i}\end{cases}$

$$\sum_{t_{1i}=y_{1i+1}}^{\infty}P_1(t_{1i},t_{2i})=\theta_2(1-\theta_2)^{t_{2i}-y_{1i}-1}(1-\theta_1-\theta_2)^{y_{1i}}$$

$$\sum_{t_{1i}=y_{1i+1}}^{\infty}P_2(t_{1i},t_{2i})=\theta_2(1-\theta_1)^{y_{1i}-t_{2i}}(1-\theta_1-\theta_2)^{t_{2i}-1}$$

- $\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty}P(t_{1i},t_{2i})=\begin{cases}\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty}P_1(t_{1i},t_{2i}) \ , if\ t_{1i}<t_{2i}\\ \sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty}P_2(t_{1i},t_{2i}) \ , if\ t_{1i}\geq t_{2i}\end{cases}$

$$\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty}P_1(t_{1i},t_{2i})=(1-\theta_2)^{y_{2i}-y_{1i}}(1-\theta_1-\theta_2)^{y_{1i}}$$

$$\sum_{t_{1i}=y_{1i+1}}^{\infty}\sum_{t_{2i}=y_{2i+1}}^{\infty}P_2(t_{1i},t_{2i})=(1-\theta_1)^{y_{1i}-y_{2i}}(1-\theta_1-\theta_2)^{y_{2i}}$$

For a Bayesian analysis, we assume the following joint prior distribution for θ_1 and θ_2 :

$$\pi(\theta_1,\theta_2)\propto\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}(1-\theta_1-\theta_2)^{\alpha_0-1},\theta_1+\theta_2<1 \qquad (15)$$

where (15) is the mass probability function of a Dirichlet $(\alpha_0,\alpha_1,\alpha_2)$ distribution with hyperparameters α_0,α_1 and α_2.

Combining the Dirichlet prior (15) distribution with the likelihood function (14), we get from the Bayes formula, the joint posterior distribution for θ_1 and θ_2 .

In the presence of a covariate vector $\mathbf{x_i}=(x_{1i},x_{2i},\dots,x_{pi})$ associated to each bivariate lifetime T_{1i} and T_{2i}, we assume logistic regression models given, respectively, by,

$$\theta_{1i} = \frac{\exp\{\boldsymbol{\beta}_1'\mathbf{x}_i\}}{1-\exp\{\boldsymbol{\beta}_1'\mathbf{x}_i\}} \text{ and } \theta_{2i} = \frac{\exp\{\boldsymbol{\beta}_2'\mathbf{x}_i\}}{1-\exp\{\boldsymbol{\beta}_2'\mathbf{x}_i\}} \qquad (16)$$

where $\boldsymbol{\beta}_\mathbf{j} = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jp})'$; $j = 1, 2$ is the vector of regression parameters i = 1,2,...,n.

2.3. Dependent lifetimes assuming a Basu-Dhar bivariate geometric distribution

In this section we consider another bivariate geometric distribution introduced in literature to analyse the bivariate lifetimes introduced in Table 1A: the Basu-Dhar (BD) bivariate geometric distribution.

The bivariate geometric distribution derived by Basu and Dhar (1995) (BD geometric distribution) has survival function is given by:

$$P(T_1 > t_1, T_2 > t_2) = p_1^{t_1}\, p_2^{t_2}\, p_{12}^{\max(t_1,t_2)} \qquad (17)$$

where 0 < p_1 < 1, 0 < p_2 < 1 and 0 < p_{12} ≤ 1. It is seen that the survival function satisfies the loss of memory property without any additional parameter restrictions, namely,

$$P(T_1 > s_1 + t, T_2 > s_2 + t \,/\, T_1 > s_1, T_2 > s_2) = P(T_1 > t, T_2 > t) = (p_1p_2p_{12})^t \quad (18)$$

The mass probability function of the BD distribution is given by,

$$P(T_1 = t_1, T_2 = t_2) = \begin{cases} (p_1)^{t_1-1}(1-p_1)(p_2p_{12})^{t_2-1}(1-p_2p_{12}) & for\ T_1 < T_2 \\ (p_1p_2p_{12})^{t_1-1}(1-p_1p_{12}-p_2p_{12}+p_1p_2p_{12}) & for\ T_1 = T_2 \\ (p_2)^{t_2-1}(1-p_2)(p_1p_{12})^{t_1-1}(1-p_1p_{12}) & for\ T_1 > T_2 \end{cases} \qquad (19)$$

The marginal distributions for T_1 and T_2 are given respectively, by,

$$P(T_1 = t_1) = P(T_1 > t_1 - 1) - P(T_1 > t_1) = (1 - p_1p_{12})(p_1p_{12})^{t_1-1}$$
$$P(T_2 = t_2) = P(T_2 > t_2 - 1) - P(T_2 > t_2) = (1 - p_2p_{12})(p_2p_{12})^{t_2-1}$$

where t_1, t_2 = 1,2,3,….. and, the means are given, respectively by,

$$E(T_1) = \sum_{t1=1}^{\infty} t_1\, P(T_1 = t_1) = (1 - p_1p_{12})^{-1}$$
$$E(T_2) = \sum_{t2=1}^{\infty} t_2\, P(T_2 = t_2) = (1 - p_2p_{12})^{-1} \qquad (21)$$

The likelihood function for p_1, p_2 and p_{12} is given by,

$$L(p_1, p_2, p_{12}) = \begin{cases} \prod_{i\in c1} P(T_{1i} = t_{1i}, T_{2i} = t_{2i}) \prod_{i\in c2} P(T_{1i} = t_{1i}, T_{2i} > t_{2i}) \\ \prod_{i\in c3} P(T_{1i} > t_{1i}, T_{2i} = t_{2i}) \prod_{i\in c4} P(T_{1i} > t_{1i}, T_{2i} > t_{2i}) \end{cases} \qquad (22)$$

where,

- $\mathrm{P}(\mathrm{T}_1 = \mathrm{t}_1, \mathrm{T}_2 = \mathrm{t}_2) = \begin{cases} (\mathrm{p}_1)^{\mathrm{t}_1-1}(1-\mathrm{p}_1)(\mathrm{p}_2\mathrm{p}_{12})^{\mathrm{t}_2-1}(1-\mathrm{p}_2\mathrm{p}_{12}) & \text{for } \mathrm{T}_1 < \mathrm{T}_2 \\ (\mathrm{p}_1\mathrm{p}_2\mathrm{p}_{12})^{\mathrm{t}_1-1}(1-\mathrm{p}_1\mathrm{p}_{12}-\mathrm{p}_2\mathrm{p}_{12}+\mathrm{p}_1\mathrm{p}_2\mathrm{p}_{12}) & \text{for } \mathrm{T}_1 = \mathrm{T}_2 \\ (\mathrm{p}_2)^{\mathrm{t}_2-1}(1-\mathrm{p}_2)(\mathrm{p}_1\mathrm{p}_{12})^{\mathrm{t}_1-1}(1-\mathrm{p}_1\mathrm{p}_{12}) & \text{for } \mathrm{T}_1 > \mathrm{T}_2 \end{cases}$

Observe that,

- $P(T_1 > t_1, T_2 > t_2) = p_1^{t_1} p_2^{t_2} p_{12}^{\max(t_1, t_2)}$
- $P(T_1 = t_1, T_2 > t_2) = \begin{cases} (p_1)^{t_1 - 1}(1 - p_1)(p_2 p_{12})^{t_2} & for\ T_1 \leq T_2 \\ (p_2)^{t_2}(p_1 p_{12})^{t_1 - 1}(1 - p_1 p_{12}) & for\ T_1 > T_2 \end{cases}$
- $P(T_1 > t_1, T_2 = t_2) = \begin{cases} (p_1)^{t_1}(p_2 p_{12})^{t_2 - 1}(1 - p_2 p_{12}) & for\ T_1 < T_2 \\ (p_2)^{t_2 - 1}(p_1 p_{12})^{t_1}(1 - p_2) & for\ T_1 \geq T_2 \end{cases}$

3. Results

In this section we present the Bayesian analysis of the dataset assuming the bivariate lifetime models introduced in section 2.

3.1. A Bayesian analysis of the survival times of Table 1A assuming the Block and Basu distribution

Assuming the Block and Basu distribution first not considering the presence of covariates, let us denote as $\delta_1$the indicator of censored observation (relapse) for T_1 and δ_2 the indicator of censored observation (death) for T_2. Assuming Gamma(1,100) prior distributions for the parameters λ_r, r = 1,2,3 of the Block and Basu bivariate geometric distribution for the lifetimes T_1 (disease-free survival time) and T_2 (overall survival time) and using the OpenBugs software (burn-in sample =10,000 and 1,000 final Gibbs samples taken every 100th sample from 100,000 simulated samples), we have in Table 1, the posterior summaries of interest. Convergence of the Gibbs sampling algorithm was verified from standard traceplots of the simulated Gibbs samples.

Table 1. Posterior summaries; Block and Basu bivariate exponential distribution

	Mean	Standard	95% Credibility interval	
		Deviation	Lower limit	Upper limit
λ_1	0.0025	0.0016	0.3830	0.0063
λ_2	0.1850	0.2510	0.0029	0.9230
λ_3	0.0075	0.0021	0.0038	0.0119
Mean 1	108.4000	22.9000	71.8800	158.1000
Mean 2	232.6000	56.4400	147.0000	358.6000
ρ_{12}	0.0013	0.0006	0.2480	0.0026
SD 1	107.9000	22.5300	71.7200	156.5000
SD 2	173.5000	44.4700	111.0000	282.7000

From the results of Table 1, we observe that the Monte Carlo estimates for the means based on the squared error loss function, there is, the posterior means for μ_1 and μ_2 (see (2)) for the disease-free survival time and the overall survival time are given, respectively, by 108.4 months and 232.6 months.

Now consider a Bayesian analysis for the bivariate survival data T_1 (disease-free survival time) and T_2 (overall survival time) of Table 1A in the presence of covariates assuming the following regression models,

$$\lambda_{\mathrm{vi}} = \alpha_{\mathrm{v}}\exp(\beta_{\mathrm{v}1}\mathrm{age}_{\mathrm{i}} + \beta_{\mathrm{v}2}\mathrm{hercep}_{\mathrm{i}} + \beta_{\mathrm{v}3}\mathrm{stage}_{\mathrm{i}} + \beta_{\mathrm{v}4}\mathrm{surgical}_{\mathrm{i}} + \beta_{\mathrm{v}5}\mathrm{pCR}_{\mathrm{i}} + \beta_{\mathrm{v}6}\mathrm{estrog}_{\mathrm{i}} + \beta_{\mathrm{v}7}\mathrm{progest}_{\mathrm{i}}) \tag{23}$$

where v=1 (disease-free survival time) and v=2 (overall survival time).

Assuming normal N(0,1) prior distributions for all regression parameters β_{1r} and β_{2r}, r = 1,2,...,7 ; α_1 ~ Gamma(1,1), α_2 ~ Gamma(1,1) and λ_3 ~ Gamma(1,100) using the OpenBugs software (burn-in sample =10,000 and 1,000 final samples taken every 50th generated sample from a Gibbs sample of size 50,000), we have in Table 2, the posterior summaries of interest.

From the results of Table 2, it is concluded that only covariate age has a significant effect (95% credible intervals for the regression parameters β_{11} and β_{21} do not include zero) for the disease-free survival times and for the overall survival times. The number of cycles of trastuzumab received before surgery does not show significant difference since zero is included in the 95% credible intervals for β_{12} and β_{22}.

Table 2. Posterior summaries of interest assuming the Block and Basu distribution in the presence of covariates

	Mean	Standard Deviation	95% Credibility interval	
			Lower limit	Upper limit
α_1	1.006	1.020	0.02049	3.7600
α_2	0.9899	1.055	0.02332	3.6040
β_{11}	-0.3400	0.1441	-0.5960	-0.0769
β_{12}	-0.04429	0.9928	-1.9200	1.8860
β_{13}	-0.04462	1.0050	-1.9430	1.9780
β_{14}	-0.01633	0.9956	-2.0180	2.0110
β_{15}	-0.03633	0.9818	-1.9000	1.8920
β_{16}	-0.09187	0.9617	-2.0250	1.7510
β_{17}	-0.1226	0.9443	-1.9460	1.6990
β_{21}	-1.1050	0.5564	-2.5100	-0.3612
β_{22}	0.0038	0.9603	-1.8740	1.8890
β_{23}	-0.1133	0.9691	-1.9890	1.7280
β_{24}	-0.0121	1.0040	-1.9660	1.8820
β_{25}	0.0690	1.0010	-1.9650	2.0110
β_{26}	0.0527	0.9837	-2.0230	1.9510
β_{27}	-0.0335	0.9877	-2.0090	1.9030
λ_3	0.0092	0.0019	0.0058	0.0133

3.2. A Bayesian analysis of the survival times of Table 1A assuming the Arnold bivariate geometric distribution

As a first analysis, we do not consider the presence of covariates assuming T_1 = disease-free survival time, T_2 = overall survival time, δ_1= censoring indicator (relapse) for T_1 and δ_2= censoring indicator (death) for T_2.

Assuming a Dirichlet(1,1,1) prior distribution with mass probability function (15) for the parameters θ_1 and θ_2 where $r = 1 - \theta_1 - \theta_2$ of the Arnold bivariate geometric distribution for the lifetimes T_1 = disease-free survival time and T_2 = overall survival time, given in Table 1A , we use the OpenBugs software (burn-in sample =10,000 e 1,000 final Gibbs samples chosen taking every 100th generated sample from a total of 100,000 Gibbs samples), to get the posterior summaries of interest (see Table 3).

Table 3. Posterior summaries of interest; Arnold bivariate geometric distribution

	Mean	Standard	95% credibility interval	
		Deviation	Lower limit	Upper limit
Mean 1	140.4000	36.2600	87.7500	222.1000
Mean2	343.6000	144.1000	160.100	720.6000
r	0.9891	0.0022	0.9845	0.9931
θ_1	0.0076	0.0018	0.0045	0.0113
θ_2	0.0033	0.0012	0.0014	0.0062

From the results of Table 3, the posterior means for the disease-free survival time and the overall survival time are given, respectively, by 140.4 months and 343.6 months, that is, similar result obtained using the Block and Basu distribution for the disease-free survival time (108.4 months) but very different for the overall survival time (232.6 months).

Assuming discrete time survival in the presence of covariates, initially let us assume independent geometric distributions for the two survival times. The geometric distribution has probability function given by:

$$P(T = t) = \theta(1 - \theta)^t \,, t = 0,1,2,3, \ldots \qquad (24)$$

where the mean is given by, $\frac{(1-\theta)}{\theta}$.

The likelihood function assuming the ith contribution is given by:

$$L_i = [P(T_i = t_i)]^{\delta_i} \, [P(T_i \geq t_i)]^{1-\delta_i} \qquad (25)$$

where $\delta_i = 1$ for a complete observation and $\delta_i = 0$ for a censored observation and $P(T_i \geq t_i) = 1 - P(T_i < t_i)$, that is, $P(T_i < t_i) = \sum_{u=0}^{t_i-1} \theta(1-\theta)^u = \theta + \theta(1-\theta) + \theta(1-\theta)^2 + \theta(1-\theta)^3 + \cdots + \theta(1-\theta)^{t_i-1}$.

From the result,

$$\sum_{k=0}^{n} ar^k = \frac{a(1-r^{n-1})}{1-r}$$ we get with $a = \theta, r = 1 - \theta$ and $n = t_i - 1$,

$$P(T_i < t_i) = \sum_{u=0}^{t_i-1} \theta(1-\theta)^u = \frac{\{\theta[1-(1-\theta)^{t_i}]\}}{\theta} = [1 - (1-\theta)^{t_i}]$$

that is,

$$P(T_i \geq t_i) = 1 - P(T_i < t_i) = 1 - [1 - (1-\theta)^{t_i}] = (1-\theta)^{t_i} \tag{26}$$

Thus, the likelihood function considering the ith contribution is given by,

$$L_i = [\theta(1-\theta)^{t_i}]^{\delta_i}\, [P(T_i \geq t_i)]^{1-\delta_i} = [\theta(1-\theta)^{t_i}]^{\delta_i}\, [\theta(1-\theta)^{t_i}]^{1-\delta_i} \tag{27}$$

In the presence of covariates, let us assume logistic regression models given by,

$$\mathrm{logit}(\theta_{\mathrm{vi}}) = \beta_{\mathrm{v0}} + \beta_{\mathrm{v1}}\mathrm{age_i} + \beta_{\mathrm{v2}}\mathrm{hercep_i} + \beta_{\mathrm{v3}}\mathrm{stage_i} + \beta_{\mathrm{v4}}\mathrm{surgical_i} + \beta_{\mathrm{v5}}\mathrm{pCR_i} + \beta_{\mathrm{v6}}\mathrm{estrog_i} + \beta_{\mathrm{v7}}\mathrm{progest_i}\,, \tag{28}$$

where v=1(disease-free survival times) and v=2 (overall survival times).

Assuming normal N(0,1) prior distributions for all regression parameters βvr, r = 0,1,2, ..., 7; v = 1,2 and using the OpenBugs software (burn-in sample = 10,000 and 1,000 samples taking every 50th generated Gibbs sample from 50,000 simulated samples), we have in Table 4 the posterior summaries of interest considering the disease-free survival times and the overall survival times. From the results of Table 4, it is possible to observe that the covariate stage has a significative effect (95% credibility interval for the corresponding regression parameters β13 and β23 do not include zero) for the disease-free survival times and for the overall survival times.

Table 4. Posterior summaries of interest; independent geometric distributions in the presence of covariates

	Mean	Standard Deviation	95% credibility interval	
			Lower limit	Upper limit
disease-free survival times				
β_{10}	-1.2000	0.8696	-2.7800	0.4977
β_{11}	-0.0152	0.0230	-0.0619	0.0291
β_{12}	-0.8317	0.6561	-2.1860	0.3560
β_{13}	-0.8810	0.3691	-1.5770	- 0.1058
β_{14}	0.1874	0.5084	-0.7282	1.2350
β_{15}	-0.6282	0.4772	-1.6170	0.2523
β_{16}	-0.3752	0.5529	-1.5060	0.7418
β_{17}	-0.5505	0.6116	-1.7720	0.6415
overall survival times				
β_{20}	-1.186	0.8751	-2.8780	0.6109
β_{21}	0.0211	0.0311	-0.0366	0.0840
β_{22}	-0.9263	0.7884	-2.5530	0.5488
β_{23}	-1.3850	0.4176	-2.1660	-0.5847
β_{24}	0.8654	0.6555	-0.3827	2.1350
β_{25}	-0.8230	0.6276	-2.0780	0.3818
β_{26}	0.0619	0.6718	-1.2790	1.3560
β_{27}	-0.6686	0.7095	-2.1230	0.6201

As a second analysis, let us consider assume the Arnold bivariate geometric distribution in presence of covariates and the regression model given by (28). In this case, we assume informative normal prior distributions using prior information (use of empirical Bayes methods, see, for example, Carlin and Louis, 2002) of Table 4 (independent geometric distributions): β10 ~ N(-1,2,1), β20 ~ N(-1.2,1), β11 ~ N(-0.015,1), β12 ~ N(-0.83,1), β13 ~ N(-0.88,1), β14 ~

N(0.18,1), β15 ~ N(-0.62,1), β16 ~ N(-0.37,1), β17 ~ N(-0.55,1), β21 ~ N(0.02,1), β22 ~ N(-0.92,1), β23 ~ N(-1.38,1), β24 ~ N(0.86,1), β25 ~ N(-0.82,1), β26 ~ N(0.06,1) and β27 ~ N(-0.66,1). Using the OpenBugs software (burn-in sample = 1,000 and 1,000 samples taking every 10th generated Gibbs sample), we have in Table 5, the posterior summaries of interest. It is important to point out, that in this case, the convergence of the Gibbs sampling algorithm using the OpenBugs software only was obtained using this class of informative priors.

From the results of Table 5, it is observed that all covariates do not have significant effect on the parameter θ1 related to the marginal distribution for the disease-free survival times (95% credibility intervals for all regression parameters include the zero value); in the same way, the covariates stage and surgery type have significant effects on the parameter θ2 related to the marginal distribution for the overall survival times (95% credibility intervals for the regression parameters β23 and β24 does not include the zero value).

Table 5. Posterior summaries of interest; Arnold bivariate geometric distribution in presence of covariates

	Mean	Standard	95% Credibility interval	
		Deviation	Lower limit	Upper limit
β_{10}	-1.8740	0.8219	-3.4810	-0.2531
β_{11}	-0.0158	0.0234	-0.0619	0.0307
β_{12}	-0.8795	0.6582	-2.1070	0.4762
β_{13}	-0.6307	0.3652	-1.3620	0.0627
β_{14}	0.2534	0.4839	-0.6496	1.2250
β_{15}	-0.7046	0.4962	-1.6820	0.2873
β_{16}	-0.3690	0.5558	-1.4590	0.6577
β_{17}	-0.6515	0.6313	-1.9100	0.5810
β_{20}	-1.7990	0.8942	-3.3820	0.0757
β_{21}	0.0233	0.0324	-0.0355	0.0868
β_{22}	-1.0570	0.7301	-2.5330	0.3101
β_{23}	-1.3260	0.4500	-2.1050	-0.4563
β_{24}	1.4880	0.7580	0.0771	3.0031
β_{25}	-0.9385	0.6561	-2.2590	0.3240
β_{26}	0.2356	0.6829	-1.0710	1.5580
β_{27}	-0.9555	0.7874	-2.5620	0.4550

3.3. A Bayesian analysis of the survival times of Table 1A assuming the Basu-Dhar bivariate geometric distribution

Assuming uniform U(0, 1) prior distributions for the parameters p_1, p_2 and p_{12} of the BD bivariate distribution geometric for the lifetimes T_1 (disease-free survival) and T_2 (overall survival) introduced in Table 1A, not considering the presence of covariates, we also have used the software OpenBugs (burn-in sample = 10,000 and a final sample of size 100; every 10th Gibbs sample), to find the posterior summaries of interest (see Table 6).

From the results of Table 6, the Monte Carlo estimates of the posterior means for the disease-free survival time and the overall survival time are given, respectively, by 111.8 months and 296.8 months, that is, similar results obtained using the Block and Basu distribution for the disease-free survival time (108.4 months) but a little different for the overall survival time (232.6 months).

Table 6. Posterior summaries of interest; bivariate BD geometric distribution with no presence of covariates

	Mean	Standard Deviation	95% credibility interval	
			Lower limit	Upper limit
Mean 1	111.8000	29.2600	69.9700	177.9000
Mean2	296.8000	121.6000	151.7000	559.6000
p_1	0.9924	0.0019	0.9882	0.9957
p_{12}	0.9981	0.0013	0.9952	0.9999
p_2	0.9981	0.0013	0.9949	0.9999

Now consider a Bayesian analysis of the discrete bivariate survival data T_1 (disease free survival) and T_2 (overall survival) in the presence of covariates with a Basu-Dhar bivariate geometric distribution and the following regression models:

$$\mathrm{logit}(p_{1i}) = \beta_{10} + \beta_{11}(\mathrm{age}_i - 48.29) + \beta_{12}\mathrm{hercep}_i + \beta_{13}\mathrm{stage}_i + \beta_{14}\mathrm{surgical}_i + \beta_{15}\mathrm{pCR}_i + \beta_{16}\mathrm{estrog}_i + \beta_{17}\mathrm{progest}_i \quad ,$$

$$\mathrm{logit}(p_{2i}) = \beta_{20} + \beta_{21}(\mathrm{age}_i - 48.29) + \beta_{22}\mathrm{hercep}_i + \beta_{23}\mathrm{stage}_i + \beta_{24}\mathrm{surgical}_i + \beta_{25}\mathrm{pCR}_i + \beta_{26}\mathrm{estrog}_i + \beta_{27}\mathrm{progest}_i \ , \quad ,$$

$$\mathrm{logit}(p_{12i}) = \beta_{30} + \beta_{31}(\mathrm{age}_i - 48.29) + \beta_{32}\mathrm{hercep}_i + \beta_{33}\mathrm{stage}_i + \beta_{34}\mathrm{surgical}_i + \beta_{35}\mathrm{pCR}_i + \beta_{36}\mathrm{estrog}_i + \beta_{37}\mathrm{progest}_i \ . \qquad (29)$$

Assuming normal N(0,1) prior distributions for all regression parameters and using the OpenBugs software (burn-in sample = 2,000 and 1,000 final samples taken every 10^{th}), we have in Table 7, the posterior summaries of interest.

From the results of Table 7, it is observed that the covariate stage has a significant effect (95% credibility interval for the regression parameters β13, β23 and β33 do not include the zero value), that is, they have significant effects at p1, p2 and p12. From the expressions (21) for the means, it is concluded that stage affects the disease free times and the overall survival times.

It is important to point out that for this model, we convergence of the MCMC simulation algorithm considering non-informative prior distributions was easily obtained with no need for informative priors as it was assumed using the Arnold bivariate geometric distribution (an advantage of the Basu-Dhar geometric distribution when compared with the Arnold geometric distribution). Additionally, we observe that the regression model assuming a Basu-Dhar geometric distribution is more sensitive to identify the significant effects of the covariates.

4. Conclusions and discussion of the obtained results

The identification of appropriated models to analyze bivariate survival data in presence of censored data and covariates is of great importance and interest to many application areas such as engineering and medicine. In the presence of a large proportion of censored data, we usually could have great difficulties to get the inferences of interest assuming standard continuous bivariate distributions introduced in the literature. In this way, the use of bivariate discrete distributions could be a good alternative to analyze bivariate lifetime data.

Table 7. Posterior summaries of interest; BD bivariate geometric distribution in the presence of covariates

	Mean	Standard Deviation	95% Credibility interval	
			Lower limit	Upper limit
β_{10}	1.3140	0.8603	-0.2721	2.9290
β_{11}	0.0138	0.0232	-0.0296	0.0606
β_{12}	0.9656	0.6458	-0.2764	2.2440
β_{13}	0.7564	0.3642	0.0151	1.4580
β_{14}	-0.1023	0.5181	-1.1240	0.8989
β_{15}	0.6145	0.4915	-0.3446	1.6110
β_{16}	0.3507	0.5735	-0.7618	1.5390
β_{17}	0.5777	0.6353	-0.6480	1.8160
β_{20}	1.1530	0.9321	-0.6813	3.0080
β_{21}	-0.0191	0.0667	-0.1559	0.1312
β_{22}	1.0900	0.8652	-0.6978	2.8240
β_{23}	1.6480	0.6862	0.6425	3.3490
β_{24}	-0.3494	0.9079	-1.9790	1.6200
β_{25}	0.7181	0.8034	-0.9457	2.2820
β_{26}	0.0945	0.8543	-1.5110	1.8830
β_{27}	0.5058	0.8394	-1.2520	2.1570
β_{30}	1.1280	0.9296	-0.7318	2.9710
β_{31}	-0.0169	0.0732	-0.1666	0.1510
β_{32}	1.0430	0.8421	-0.5590	2.6840
β_{33}	1.7460	0.6663	0.6496	3.3100
β_{34}	-0.2448	0.9106	-1.9580	1.6260
β_{35}	0.6825	0.8014	-0.9267	2.2660
β_{36}	0.1583	0.8724	-1.5520	1.9140
β_{37}	0.4331	0.8699	-1.3040	2.1410

The use of Bayesian methods and MCMC simulation techniques also open a new horizon in the analysis of such data, as noted in the application data of breast cancer introduced in this article. Moreover, we observed that the use of bivariate models considering discrete data can be more sensitive and efficient in getting the inferences of interest.

Important inferences were obtained for our application considering the breast cancer data introduced in Table 1A.

Assuming the three bivariate lifetime distributions, we see that there are no significant differences between the survival times for the patients receiving at least four or less than four cycles of trastuzumab before surgery.

The factors age, stage and surgical has significant effect on the survival times of the patients (not at the same time). Under a regression model for the parameters of the Block and Basu distribution, we see that only the covariate age has significant effect (zero not included in the 95% credible intervals for β_{11} and β_{21} in the regression model (23)).

Under a regression model for the parameters of the Arnold distribution assuming non-informative priors and independent lifetimes T_1 and T_2, we see that the covariate stage has significant effect (zero not included in the 95% credible intervals for the associated regression parameters in model (28)) and assuming informative priors with the bivariate Arnold model (T_1 and T_2 dependent lifetimes), we see that the covariate stage and surgery has significant effect.

Finally, under a regression model for the parameters of the Basu-Dhar bivariate distribution, only the covariate stage has significant effect (zero not included in the 95% credible intervals for the associated regression parameters in model (29)).

From these inference results considering the three different models and medical interpretations of cancer experts, we have some comments:

Incidence rates of breast cancer increase rapidly up to 50 years old. After this age, the increase occurs more slowly, which strengthens the participation of female hormones in the etiology of the disease. However, breast cancer seen in young women have clinical and epidemiological characteristics very different from those in older women. They are generally more aggressive, have a high mutation rate of the presence of the BRCA1 and BRCA2 genes in addition to overexpress the gene for human epidermal growth factor receptor 2 (HER-2).

Patients with advanced stages also in general have more relapses and die more as it is observed on the plots of the Kaplan and Meier non-parametric estimates considering the stages II and III (not included to save space).

The type of surgery is a confounding factor. Radical surgery does not directly affect the survival times, actually patients who undergo radical surgery are usually in the stage 3. The stage 2 patients are subjected to more conservative surgery.

The Monte Carlo estimates for the posterior means of μ_1 and μ_2 (means of T_1 - disease free survival and T_2 - overall survival) are very similar assuming the Block and Basu bivariate exponential distribution (108.4 and 232.6 months) and the Basu-Dhar distribution (111.8 and 296.8 months), but the 95% credible intervals are different (larger for the Basu-Dhar distribution). Thus, the Monte Carlo estimates for the posterior means of μ_1 and μ_2 are very different assuming the Arnold bivariate exponential distribution (140.4 and 343.6 months) with larger 95% credible intervals.

It is important to point out, that further discrimination methods should be developed for the the comparison of the different bivariate lifetime models assumed in the analysis of the breast cancer data of Table 1A in presence of a great number of censored observations, where some usual existing discrimination methods as the DIC (Deviance Information Criterion) introduced by Spiegelhalter et al (2002) could be not reliable to discriminate the proposed models.

As an empirical way to compare the proposed models, we could compare the obtained Monte Carlo estimates of the posterior means for the disease-free survival times and the overall survival times with a non-parametrical estimate (Kaplan-Meier estimates). From the estimates given in Table 8, we observe that the Block and Basu estimates are more close to the Kaplan-Meier estimates for the means, a possible indication of better fit by the data. Observe that the data set presents a great proportion of censored observations, a possible indication of immune or cured individuals and the proposed bivariate models are more sensible to capture this fact. Other possibility in a future work: use of bivariate cure fraction models.

Finally, it is interesting to observe that using the popular Weibull distribution (see for example, Abrams et al, 1996) with two parameters, the Monte Carlo estimate of the posterior mean for the disease-free survival time assuming non-informative priors is given by 81.71 and for the overall survival time is given by 139.9. In this case, the maximum likelihood estimates for the means are given, respectively by 73.0497 and 96.6122, but the Weibull regression models were not able to detect any covariate effect.

Table 8. Estimates for the means of the disease-free survival times and the overall survival times

Estimate	disease-free survival times	overall survival times
Kaplan-Meier	63.0	73.5
Block and Basu	108.4	232.6
Arnold	140.4	343.6
Basu-Dhar	111.8	296.8

Acknowledgements

This research was carried out with support of a brazilian grant of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil). The authors acknowledge the suggestions from a reviewer and the editor which improved the presentation of the manuscript.

References

[1] Abrams, K.; Ashby, D. and Errington, D.A (1996). Bayesian approach to Weibull survival models-application to cancer clinical trial, *Lifetime Data Analysis*, 2(2), 159-174.

[2] Achcar, J.A. and Leandro, R.A. (1998). Use of Markov Chain Monte Carlo methods in a Bayesian analysis of the Block and Basu bivariate exponential distribution, Annals of the Institute of Statistical Mathematics, 50 , 403-416.

[3] Achcar, J.A. and Santos, C.A. (2011). A Bayesian Analysis for the Block and Basu Bivariate Exponential Distribution in the Presence of Covariates and Censored Data, Journal of Applied Statistics, 38 , 2213-2223.

[4] Arnold, B.C. (1975). A characterisation of the exponential distribution by multivariate geometric compounding, Sankhya, Series A, 37, 164-173.

[5] Arnold, B.C. and Strauss, D. (1988). Bivariate distributions with exponential conditionals. Journal of the American Statistical Association, 83(402), 522–527.

[6] Basu, A. P. and Dhar, S. (1995). Bivariate geometric distribution, Journal Applied Statistical Science, 2(1), 33-44.

[7] Block, H.W. and Basu, A.P. (1974). A continuous bivariate exponential extension, Journal of the American Statistical Association,69, 1031-1037.

[8] Brody, J.G. et al. (2007). Environmental pollutants, diet, physical activity, body size, and breast cancer: where do we stand in research to identify opportunities for prevention?, Cancer , 109(12), 2627–34.

[9] Carlin, B. P. and Louis, T. (2002). A Bayes and Empirical Bayes Methods for Data Analysis. Chapman Hall.

[10] Carroll, K.J. (2003). On the use and utility of the Weibull model in the analysis of survival data, Contemporary Clinical Trials, 24(6), 682-701.

[11] Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler, The American Statistician, 46, 167–174.

[12] Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, The American Statistician, 49, 327–335.

[13] Coelho-Barros, E.; Achcar, J.A. and Mazucheli, J. (2016). Bivariate Weibull Distributions Derived From Copula Functions In The Presence Of Cure Fraction And Censored Data, *Journal of Data Science,***14**, 295-316.

[14] Colditz, G, A.; Kaphingst, K.A.; Hankinson, S. E. and Rosner, B. (2012). Family history and risk of breast cancer: nurses' health study, *Breast Cancer Research and Treatment*, **133**(3), 1097–1104.

[15] Cox, D. R. (1972). Regression models and life tables, Journal of the Royal Statistical Society, B, 34, 187–220.

[16] Cuzick, J. et al. (2013). Selective estrogen receptor modulators in prevention of breast cancer: an updated meta-analysis of individual participant data, Lancet, 381(9880), 1827–34.

[17] Davarzani, N.; Achcar, J.A., Smirnov, E.N. and Peeters, R. (2015). Bivariate lifetime geometric distribution in presence of cure fractions, *Journal of Data Science,* **13**, 755-770.

[18] DeSantis, C.; Ma, J.; Bryan, L. and Jemal, A. (2014). Breast cancer statistics: 2013, *CA: A Cancer Journal for Clinicians,* **64**(1), 52-62.

[19] Downton, F. (1970). Bivariate exponential distributions in reliability theory, Journal of the Royal Statistical Society, B, 32, 408–417.

[20] Ferro, R. (2012). Pesticides and Breast Cancer, *Advances in Breast Cancer Research*, **1**(3), 30–35.
[21] Freund, J. E. (1961). A bivariate extension of the exponential distribution, Journal of the American Statistical Association, 56, 971–977.

[22] Gaffield, M.E.; Culwell, K.R. and Ravi, A. (2009). Oral contraceptives and family history of breast cancer, Contraception , 80(4), 372–80.

[23] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities, Journal of the American Statistical Association, 85, 398–409.

[24] Hendrik, R.E. (2010). Radiation doses and cancer risks from breast imaging studies, Radiology, 257(1), 246–53.

[25] Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions, Biometrika, 3(73), 387–396.

[26] Kathlenborn, C.; Modugno, F.; Potter, D.M. and Severs, W.B. (2006). Oral contraceptive use as a risk factor for premenopausal breast cancer: a meta-analysis, *Mayo Clinic proceedings. Mayo Clinic*, **81**(10),1290–302.

[27] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations , Journal of the American Statistical Association. , **53**(282), 457–481.

[28] Kelsey, J.L. (1993). Breast cancer epidemiology: summary and future directions, Epidemiology Review,15(1), 256-63.

[29] Lacroix, M. (2006). Significance, detection and markers of disseminated breast cancer cells, Endocrine-Related Cancer (Bioscientifica), 13(4),1033–67.

[30] Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data, John Wiley & Sons, New York.

[31] Lee, E.T. and Wenyuwang, J. (2003). Statistical methods for survival data analysis, third edition, John Wiley & Sons, New Jersey.

[32] Louzada-Neto, F.; Suzuki, A.K., A.; Cancho, V. G.; Prince, F.L. and Pereira, G.A. (2012). The Long-Term Bivariate Survival FGM Copula Model: An Application to a Brazilian HIV Data, *Journal of Data Science*, **10**, 511-535.

[33] Marshall, A.W. and Olkin, I. (1967). A generalized bivariate exponential distribution, *Journal* of Applied Probability, 4, 291-302.
[34] McPherson, K.; Steel, C.M. and Dixon, J.M. (2000). ABC of breast diseases: breast cancer-epidemiology, risk factors, and genetics, BMJ, 321(7261), 624-628.

[35] Nair, K.R.M. and Nair, N.U. (1988). On characterizing a bivariate geometric distribution, Journal of the Indian Statistical Association, 26, 45-49.

[36] Nelson, H.D. et al. (2012). Risk factors for breast cancer for women aged 40 to 49 years: a systematic review and meta-analysis, *Annals of Internal Medicine*, **156**(9), 635–48.

[37] Reeder, J.G. and Vogel, V.G. (2008). Breast cancer prevention, *Cancer treatment and research* , **141**,149–64.

[38] Saunders, C. and Jassal. S. (2009). Breast cancer (1. ed.). Oxford: Oxford University Press. Chapter 13.

[39] Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer, *New England Journal of Medicine*, **360**(8),790–800.

[40] Spiegelhalter, D.J.; Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, Journal of the Royal Statistical Society, Series B, 64(4), 583-639.

[41] Spiegelhalter, D.J.; Thomas, A.; Best, N. G. and Gilks, W. R. (2003). WinBUGS User Manual (version 1.4), MRC Biostatistics Unit, Cambridge, UK.

[42] Tiezzi, D.G. (2009). Epidemiologia do câncer de mama, Revista Brasileira de Ginecologia e Obstetricia, 31(5), 213-215.

[43] World Cancer Report (2014a). World Health Organization, Chapter 5.2.

[44] World Cancer Report (2014b). International Agency for Research on Cancer, World Health Organization. ISBN 978-92-832-0432-9.

[45] Wu, A.H.; Yu, M.C.; Tseng, C.C. and Pike, M.C. (2008). Epidemiology of soy exposures and breast cancer risk, *British Journal of Cancer,* **98**(1), 9–14.

[46] Yager, J.D. and Davidson, N.E. (2006). Estrogen carcinogenesis in breast cancer, *New England Journal of Medicine*, 354(3), 270–82.

[47] Yang,L. and Jacobsen, K.H. (2008). A systematic review of the association between breastfeeding and breast cancer, Journal of Women's Health, **17**(10), 1635–1645.

Appendix A - Data set

Table 1A. Data of 54 patients with breast cancer

Ident	Age	Hercep	Stage	Surgical	pCR	Estr	Proge	Relapse	DFS	Death	TS
2	50	1	3	1	1	0	0	0	60	0	60
8	24	1	2	0	1	1	1	0	45	0	45
9	44	1	3	1	1	1	1	0	83	0	83
11	43	1	3	1	1	1	1	0	53	0	53
14	29	1	3	0	0	1	1	1	30	0	58
16	40	1	3	1	0	0	0	0	72	0	72
18	48	1	3	1	1	1	1	0	30	0	30
19	62	1	3	0	0	0	0	0	65	0	65
20	51	1	3	0	1	0	0	0	68	0	68
21	48	1	3	1	1	0	0	1	20	0	60
22	50	1	3	0	0	1	0	0	64	0	64
25	44	1	3	1	0	0	1	0	33	0	33
29	63	1	2	1	1	0	0	0	37	0	37
30	52	1	3	1	0	1	0	0	27	0	27
32	35	1	3	1	0	0	0	1	22	0	59
33	41	1	3	1	0	0	0	0	34	0	34
35	57	1	3	1	1	0	0	0	34	0	34
40	57	1	3	0	1	1	0	0	46	0	46
41	32	1	3	1	0	0	0	0	32	0	32
42	71	1	3	1	0	0	0	0	66	0	66
49	37	1	3	1	0	1	0	1	53	1	61
50	62	1	2	0	1	0	0	0	55	0	55
51	42	1	3	1	1	0	0	0	65	0	65
52	30	1	3	1	0	1	1	1	44	1	50
53	60	1	3	1	0	1	0	1	15	1	53
54	51	1	3	1	0	0	0	1	33	1	37
55	62	1	3	1	0	0	0	1	12	1	17
57	47	1	2	0	0	1	1	0	59	0	59
58	42	1	3	1	0	1	0	0	39	0	39
60	42	1	3	1	0	1	1	0	29	0	29
63	63	1	3	1	0	0	0	0	35	0	35
65	57	1	3	1	1	0	0	1	22	1	28
66	56	1	2	*	*	0	0	*	8	1	8
68	63	1	3	1	0	0	0	*	22	0	22
70	30	1	3	0	0	0	0	1	47	0	62
71	34	1	2	1	0	0	0	0	25	0	25
72	39	1	3	1	1	0	0	1	48	0	58
73	41	1	3	1	0	1	1	1	49	0	83
74	58	1	2	0	1	0	0	1	31	0	41
3	57	2	3	0	0	0	0	0	42	0	42
4	39	2	3	0	1	0	0	0	30	0	30
7	65	2	3	0	0	1	1	0	30	0	30
10	54	2	3	1	0	1	1	0	56	0	56
13	53	2	3	1	1	0	0	0	32	0	32

23	49	2	3	0	0	1	1	0	40	0	40
26	57	2	3	1	1	0	0	1	39	1	44
27	41	2	3	1	0	1	1	0	37	0	37
31	62	2	3	0	0	0	0	1	24	0	34
37	56	2	3	0	1	1	0	0	58	0	58
39	52	2	2	1	1	1	1	0	29	0	29
43	49	2	3	1	1	0	0	0	44	0	44
48	40	2	3	1	1	1	1	0	22	0	22
61	51	2	2	1	1	1	1	0	31	0	31
75	48	2	2	0	1	1	1	0	16	0	16

Received August xxxx; accepted xxxx

Tatiana Reis Icuma
Department of Social Medicine, Medical School
University of São Paulo
Ribeirão Preto, SP, Brazil
tati.icuma@usp.br

Isabela Panzeri Carlotti Buzatto
Department of Obstetrics and Gynecology, Medical School
University of São Paulo
Ribeirão Preto, SP, Brazil
isabelacarlotti@yahoo.com.br

Daniel Guimarães Tiezzi
Department of Obstetrics and Gynecology, Medical School
University of São Paulo
Ribeirão Preto, SP, Brazil.
dtiezzi@fmrp.usp.br

*Jorge Alberto Achcar (corresponding author)
Department of Social Medicine, Medical School
University of São Paulo
Ribeirão Preto, SP, Brazil.
achcar@fmrp.usp.br
Tel: +55 16-997824296

Nasser Davarzani
Department of Knowledge Engineering
Maastricht, the Netherlands.
n.davarzani@maastrichtuniversity.nl