A BAYESIAN WEIBULL ANALYSIS OF BREAST CANCER DATA WITH LONG-TERM SURVIVORS IN PARANA STATE, BRAZIL

Talita Evelin Nabarrete Tristão de MORAES[1]
Isolde Terezinha Santos PREVIDELLI[1]
Giovani Loiola da SILVA[2]

■ ABSTRACT: Breast cancer is one of the most common diseases among women worldwide with about 25% of new cases each year. In Brazil, 59,700 new cases of breast cancer were expected in 2019, according to the Brazilian National Cancer Institute (INCA). Survival analysis has been an useful tool for the identifying the risk and prognostic factors for cancer patients. This work aims to characterize the prognostic value of demographic, clinical and pathological variables in relation to the survival time of 2,092 patients diagnosed with breast cancer in Parana State, Brazil, from 2004 to 2016. In this sense, we propose a Bayesian analysis of survival data with long-term survivors by using Weibull regression models through integrated nested Laplace approximations (INLA). The results point to a proportion of long-term survivors around 57.6% in the population under study. In regard to potential risk factors, we namely concluded that 40-50 year age group has superior survival than younger and older age groups, white women have higher breast cancer risk than other races, and marital status decreases that risk. Caution on the general use of these results is nevertheless advised, since we have analyzed population-based breast cancer data without proper monitoring by a health professional.

■ KEYWORDS: Weibull regression; Cure fraction; INLA; Bayesian statistics; Survival analysis.

[1]Universidade Estadual de Maringá - UEM, Programa de Pós Graduação em Bioestatística - Departamento de Estatística, CEP: 87020-790, Maringá, PR, Brazil. E-mail: *talita_evel@hotmail.com; isoldeprevidelli@gmail.com*
[2]Universidade de Lisboa, Centro de Estatística e Aplicações (CEAUL) - FCUL & Dep. Matemática - IST, CEP:1049-001, Lisboa, Portugal. E-mail: *giovani.silva@tecnico.ulisboa.pt.*

1 Introduction

Cancer treatment and prevention have made progress over the past few decades and continue to have the attention of medical researchers and related fields. As reported by the Brazilian National Cancer Institute (INCA, 2019), the estimated incidence of breast cancer in 2019 in Brazil was 59,700 new cases, representing 29.5% of the various types of cancer in women. Araújo and Fernandes (2008) discuss the impact of the diagnosis and the confrontation of the breast cancer such as psychological effects caused on the personal image and feelings of fear of death since it is an irremediable illness. Among the socioeconomic, demographic, behavioral, regional and health factors interrelated to breast cancer prevention, Rodrigues *et al.* (2015) pointed out that women with a good socioeconomic status are the profile of women who are best at preventing breast cancer. This type of cancer occurs mostly in women, not exempting males but with a lower incidence, and represents approximately 1% of total cases of the disease (APARÍCIO, 2013).

According to INCA (2019), breast cancer can evolve in different ways because the disordered multiplication of breast cells that can invade adjacent tissues or other more distant body organs. These behaviors are due to the distinct characteristic of each tumor. It should also be noted that there are several types of breast cancer, defined by the International Classification of Diseases (ICD) for Oncology. Breast cancer ICD is an useful tool that allows more information about the pathologies cataloged by the World Health Organization (WHO), ranging from C500 to C509. Sometimes there is a need to study the effect of these DCI groups separately, investigating the individual behavior related to the location of each tumor.

Several survival models have been used to analyze cancer data such as parametric survival models e.g. Weibull regression (see details in LEE and WANG, 2003; LAWLESS, 2003; COLOSIMO and GIOLO, 2006) or semiparametric survival models e.g. Cox regression (COX, 1972). These regression models, also known by proportional hazards models, are very popular in the survival analysis due to their easy implementation and interpretation. Alternatively, other modeling approaches can be employed for cancer data, for instance, Suguiura (2017) presented a study of breast cancer that took into account the geographic information of the Health Units in the State of Parana under a multilevel Multinomial model with Gaussian random effects. Its main objective was to verify the frequency of occurrence of a certain type of ICD attending the patient's residence local, apart from some cancer risk factors.

In some cancer clinical trials, there has been a significant increase in the proportion of patients who do not experience the event of interest during the study period, such patients being commonly known as long-term survivors. The survival models that contemplate patients 'immune' to the event of interest are called survival models with long-term survivors or with 'cure' fraction (see e.g. MALLER and ZHOU, 1996; IBRAHIM *et al.*, 2001; LAMBERT, 2007). These models have become the target of interest on the part of health managers, who want to keep up with trends in the survival of patients with 'curable' diseases.

Based on an experience with calculated survivorships of patients following treatment for cancer, Berkson and Gage (1952) developed a simple function, in terms of two physically significant parameters, which fits such survivorship data very well. These two parameters can be used to briefly compare the mortality of two differently treated groups, types of cancer or other characteristics. One parameter represents the proportion of the population that is subject only to mortality rates 'normal' ('cured'), while the other is the cancer mortality rate, for which the rest of the population ('uncured') is subject. Also for cancer studies, Lambert *et al.* (2006) report that cure occurs when the mortality rate (risk) returns at the same level as expected in the general population.

Some studies have been published on survival models with cure fraction, incorporating different forms of cure fraction. For example, Sposto (2002) analyzed data from children with lymphomas and leukemias, with a significant proportion of patients cured after therapy through a parametric non-mixing model to incorporate mortality and thus obtain the estimates of the cure fraction. Cucchetti *et al.* (2015) performed a survival analysis with a cure fraction for patients after hepatectomy of colorectal liver metastases. Ramires *et al.* (2018) have used generalized additive models to investigate the proportion of cure rate in women diagnosed with breast cancer. For further details on survival models with cure fraction, see the following works and their references: Yakovlev and Tsodikov (1996) for biologically significant inferences from cancer data, Ibrahim *et al.* (2001) for cure models from a Bayesian perspective, Achcar *et al.* (2012) and Martinez *et al.* (2013) for using mixture and non-mixture cure fraction models, and Amico and van Keilegom (2018) who do a review on cure survival models.

This work aims to characterize the prognostic value of demographic, clinical and pathological variables in relation to the survival time of breast cancer patients from a population-based data, obtained through INCA and involving 2,092 patients who were observed with breast cancer in the Parana State, Brazil, from 2004 to 2016. In this sense, we propose a Bayesian analysis of survival data with long-term survivors by using Weibull regression models through integrated nested Laplace approximations (INLA), which is a method for approximate Bayesian inference and encompasses a large family of models that are used in practice. INLA has been promoted as a fast alternative to MCMC namely for disease mapping applications (CARROL *et al.*, 2016; DE SMEDT *et al.*, 2015). The article is organized as follows: Section 2 presents the methodology used in survival analysis with fraction of long-term survivors, while Section 3 exposes both descriptive analysis of the real breast cancer data and the inferential results of the proposed models using a Bayesian approach via INLA. Some conclusions are presented on the proposed survival analysis, including a criticism about the limitations found both in the analyzed data set and in the chosen software. Ultimately, a simple example of INLA code for fitting Weibull survival model with long-term survivors is shown in Appendix.

2 Methods

2.1 Material

The breast cancer data that motivated this work were obtained from INCA (*Instituto Nacional de Câncer José Alencar Gomes da Silva*), which is the body of the Ministry of Health responsible for the development and coordination of integrated actions in the prevention and control of cancer in Brazil. Those actions include hospital care provided directly and free of charge to cancer patients within Brazil's Unified Public Health System (SUS), and interventions in other strategic areas, such as prevention and early detection, training, research and epidemiological information (INCA, 2019).

The data set reports population-based information of female patients diagnosed with breast cancer in a reference hospital in the Parana State, Brazil, from 2004 to 2008, and observed in treatment until 2016. This year choice is mainly aimed at ensuring sufficient information to validate cure fraction models. So, this cross-sectional retrospective study involves $2,234$ patients for whom some demographic, clinical and pathological variables were collected. However, due to presence of missing observations of some variables and the occurrence of death dates equal to diagnosis dates, the total number of patients under study was reduced to $2,092$ patients.

Some of these variables, which are already known as risk factors for patients with breast cancer, were unfortunately not disclosed in the data set such as treatment type and disease stage. On the other hand, this has become a challenge for us as we invest in the search for risk factors not completely studied in the breast cancer literature, e.g., marital status and race. Note that we exclude male patients since the simultaneous study of male and female is controversial in this sort of cancer and, in addition, their number of cases is small.

Taking into account our initial objectives, we defined our response variable hereinafter called survival time. That is, the time between the date of diagnosis and the date of death due to breast cancer. For patients who did not die of breast cancer by the end of 2016, their survival times were calculated replacing the date of death by the date of the last hospital contact or of death due to another disease. For convenience, survival times were represented in years, whereas covariate Age was categorized into three Age group: less than forty years, between 40 and 50 years, and more than 50 years (MCGUIRE *et al.*, 2015), being the latter also associated with the menopause period (RODRIGUES *et al.*, 2015). In Section 3.1, we come back this matter doing a preliminary data analysis for all these variables.

2.2 Models with long-term survivors

Survival analysis deals with survival times of individuals observed during a given period, e.g., the elapsed time between the study entry (cancer diagnosis) and the occurrence of the event of interest (breast cancer death) for each patient. One of the main characteristics of survival data is the presence of censoring i.e. the event

does not occur for some individuals under study. That can be frequent in cohort studies and clinical cancer trials (YAKOVLEV and TSODIKOV, 1996).

Survival data set with n patients is usually represented by the pair (t_i, δ_i), where t_i is the survival time of the patient i and δ_i its death indicator, i.e. $\delta_i = 1$, if patient i has the event of interest, or $\delta_i = 0$, otherwise (censoring), $i = 1, \ldots, n$. A probabilistic distribution is often adopted for modeling survival times (T) that are characterized by its survival function, $S(t)$ or hazard function $h(t)$, $t \geq 0$. The former is the probability that the patient survives at least until the moment t, while the latter is the rate of occurrence of death at the moment t. In fact, if one knows the form of $S(t)$, one can derive the corresponding $h(t)$, and vice versa, by the relationship between the two ones, i.e.,

$$S(t) \equiv P(T \geq t) = \frac{f(t)}{h(t)}, \tag{1}$$

where $f(t)$ is the probability density function of the observable survival times that are here considered absolutely continuous random variables. For further details on survival analysis, see e.g. Lawless (2003), Lee and Wang (2003), and Colosimo and Giolo (2006).

Survival models have been employed under parametric, non-parametric and semi-parametric approaches considering the data specificity. For example, the observation of individuals who are not susceptible to the occurrence of the event of interest during the study period. In this case, parametric survival models can be a good choice for modeling both the survival times and the proportion of long-term survivors or fraction of cure. Survival models with cure fraction were initially proposed by Boag (1949), and later the methodology of cure proportion was introduced by Berkson and Gage (1952). These used the survival function in the form of mixture, where the population is divided into two parts: one that represents survival time for 'uncured' individuals and the other involving a distribution for survival time for 'cured' individuals.

According to Ibrahim *et al.* (2001), the cure (rate) model has been used to analyze survival data for various types of cancer such as breast cancer, leukemia, prostate cancer, head and neck cancer, where a significant proportion of patients are 'cured' after sufficient follow-up. In this model, a certain fraction p is considered for the cured population, whereas the remaining fraction $1 - p$ is for the uncured population. Therefore, the survival function for the entire population is given by

$$S(t) = p + (1 - p)\, S_0(t) = \frac{(1 - p)\, f_0(t)}{h(t)} \tag{2}$$

where $S_0(t)$ and $f_0(t)$ are respectively the (proper) survival and density functions for individuals who are at death risk (non-cured group), and $h(t)$ is the hazard function for the entire population, $t > 0$, $0 < p < 1$. Note that $\lim_{t \to \infty} S(t) = p$ in (2) and since $p \neq 0$, $S(t)$ is not a proper survival function, where p is the proportion of long-term survivors i.e. the fraction of individuals not susceptible to the event of interest (see details in MALLER and ZHOU, 1996; IBRAHIM *et al.* 2001).

Consequently, for n individuals who constitute a (cancer) data set, denoted here by $\mathcal{D} = \{(t_i, \delta_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$, one can assume a probability model indexed by a parameter $\boldsymbol{\theta}$ and therefore the corresponding likelihood function is expressed as

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} [(1-p) f_0(t_i|\boldsymbol{\theta})]^{\delta_i} [p + (1-p) S_0(t_i|\boldsymbol{\theta})]^{1-\delta_i}, \tag{3}$$

where t_i is the survival time, δ_i is the death indicator, and $\mathbf{x}_i$ is the observed covariates vector for the individual i. Assuming the Weibull model $\mathcal{W}(\gamma, \lambda)$, with shape parameter γ and scale parameter λ, the survival and density functions in (3) are equal to $S_0(t|\gamma, \lambda) = \exp(-\lambda t^{\gamma})$ and $f_0(t|\gamma, \lambda) = \gamma \lambda t^{\gamma - 1} \exp(-\lambda t^{\gamma})$, respectively. Note that the covariates may be introduced into the Weibull model using the parameter $\lambda \equiv \exp(\mathbf{x}^T \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the regression coefficient vector associated with a generic covariate vector $\mathbf{x}$.

Under a Bayesian perspective, inference on the parameters of the Weibull cure survival model, denoted here as $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}, p)$, will be made from the joint posterior distribution that is represented in a more simplified way as

$$\text{Posterior distribution} \ \propto\ \text{Likelihood function} \times \text{Prior distribution}, \tag{4}$$

where the likelihood function is the equation (3) related to Weibull model mentioned above, and prior distribution refers to the prior information of the model parameters. For further details about the construction of posterior distributions (4), see e.g. Paulino *et al.* (2018) or Amaral-Turkman *et al.* (2019).

Even considering well-known probability distributions for composing the likelihood and prior distributions, the posterior distribution (4) may not be simple to work with. In this sense, Markov chain Monte Carlo (MCMC) methods are widely used for making inference from a Bayesian perspective, whereas integrated nested Laplace approximations (INLA) have became more efficient in last years. Our choice fell on the latter due to the fact that the proposed cure survival model is more easily implemented under INLA approach. In fact, the package R-INLA (RUE *et al.*, 2009) was employed as a result of its better performance in executing such survival models. In addition, the software R (2020) was also used to obtain other calculations and plots presented here.

Finally, an important part in any statistical analysis is the evaluation, selection and comparison of models. So, we need to define criteria that allow e.g. the selection of a parsimonious model. In other words, a model capable of satisfactorily and simply explaining survival times in relation to explanatory variables. In the Bayesian context, several measures have already been proposed such as predictive performance measures or predictive Bayesian leave-one-out residuals. For instance, Conditional Predictive Ordinate (CPO) that is based on leave-one-out cross validation idea. The most suitable model is the one that has the largest sum of the logarithms of the individual conditional predictive ordinates, i.e. $CPO = \sum_{i=1}^{n} \log CPO_i,\ i = 1, \ldots, n$ (see PAULINO et al, 2018). Other predictive performance measures are *Deviance Information Criterion* (DIC) and

Watanabe-Akaike Information Criterion (WAIC), proposed by Spiegelhalter *et al.* (2002) and Watanabe (2010), respectively. Small values of these two measures indicate the best fitted model. For some details of diagnostic measures and adequacy of Bayesian models, see e.g. Paulino *et al.* (2018) and Wang *et al.* (2018).

3 Results

3.1 Preliminary analysis

The data set here refers to patients observed with breast cancer in Parana State, Brazil, from 2004 to 2016. In a preliminary phase, the covariates 'patient's residence city' and 'disease code/description' were excluded as result of the presence of too many missing values and consequently this would bring little information to the study. During the study, 11 male patients were also excluded from the study as long as the objective here is to study female patients with breast cancer. Thus, the final data set has $2,092$ patients and several explanatory variables.

Table 1 shows the descriptive statistics of the categorized variables, as well as the definition of their categories, thus describing the characteristics of breast cancer patients. It can be noted that most patients are over 50 years (62.86%), and 73.42% and 46.27% of them are white and married, respectively. Regarding to morphology, ductal Lobular has 65.58% of patients, while ductal carcinoma has 9.85%, corroborating with the breast cancer literature. One intriguing result is to have $1,468$ patients who were alive at the end of the study, considered a high percentage of long-term survivors (70.17%). Censoring was defined here for patients who were alive or died of other causes excluding breast cancer (96 patients).

A potential relevant covariate for cancer is topography, represented by the ICD codes that are responsible for classifying the type of cancer, allowing for more information on the pathologies, cataloged by WHO. ICD is termed by the following codes: C500 - Malignant neoplasm of nipple and areola of the breast; C501 - Malignant neoplasm of central portion of the breast; C502 - Malignant neoplasm of upper-inner quadrant of the breast; C503 - Malignant neoplasm of lower-inner quadrant of the breast; C504 - Malignant neoplasm of upper-outer quadrant of the breast; C505 - Malignant neoplasm of lower-outer quadrant of the breast; C506 - Malignant neoplasm of the axillary tail of the breast; C508 - Malignant neoplasm of overlapping sites of the breast; C509 - Malignant neoplasm of unspecified site of the breast. By reason of low frequency of some ICD categories, we opted for grouping ICD as follows: C504, C509 and the other ICD (C500, C501, C502, C503, C505, C506, and C508).

From Table 1, we also note that 57.79% and 23.57% of the patients were diagnosed with unspecified malignant breast neoplasia and malignant neoplasm of upper-outer quadrant of the breast, respectively. The excess of C509 cases in topography (first percentage) is still a subject that deserves further study, especially with the potential difficulty of physicians in filling this information in the patient's record.

Table 1 - Characteristics of breast cancer patients in Parana State, Brazil, observed from 2004 to 2016

Variable	Description	Total	%	Censoring(%)	Death(%)
Age group	< 40	226	10.80	7.84	2.96
	$40 - 50$	551	26.34	21.46	4.88
	> 50	1,315	62.86	45.46	17.40
Race	White	1,536	73.42	51.00	22.42
	Other	556	26.58	23.76	2.82
Marital status	Married	968	46.27	36.42	9.85
	Other	1,124	53.73	38.34	15.39
Topography	C509	1,209	57.79	42.45	15.34
(ICD)	C504	493	23.57	18.26	5.31
	Other	390	18.64	14.05	4.59
Morphology	Ductal Lobular	1,372	65.58	48.61	16.97
	Ductal Carcinoma	206	9.85	7.41	2.44
	Other	514	24.57	18.74	5.83
Status	Breast cancer death	528	25.24	-	25.24
	Alive & other deaths	1,564	74.76	74.76	-
	Total	2,092			

The Figure 1 presents Kaplan-Meier curves for the survival times, including their 95% confidence intervals, for all breast cancer patients (Figure 1-A) and for each categorized covariate in study. Graphically, there is a stabilization of the empirical survival function for all patients, reaching a plateau around 0.75. Women with malignant neoplasm of upper-outer quadrant of the breast (C504) tend to survive more than women with other ICD (Figure 1-B), and note that the C509 and other ICD survival curves intersect a few times over time. Patients who claim to be white have lower breast cancer survival than non-white patients (Figure 1-C). There are no difference between the survival curves of patients who have different categories of morphology (Figure 1-D). Married patients survive more than those who are unmarried (Figure 1-E), while patients between 40 and 50 years old are more likely to survive breast cancer than patients with 40 years old and over 50 years old (Figure 1-F).

We also note from Figure 1 that the survival functions estimated by the Kaplan-Meier estimator do not tend to zero as survival time tends to infinity. This indicates that breast cancer patients in study may have a relevant proportion of long-term survivors and consequently, for the patients not susceptible to the event of interest (death due to breast cancer), the survival model with cure fraction (2) may be more appropriate.

To conclude the preliminary analysis, some survival distributions were informally fitted to the survival times of breast cancer data, and the results pointed out empirical evidence that the Weibull distribution was the best fitted distribution related to these survival data.

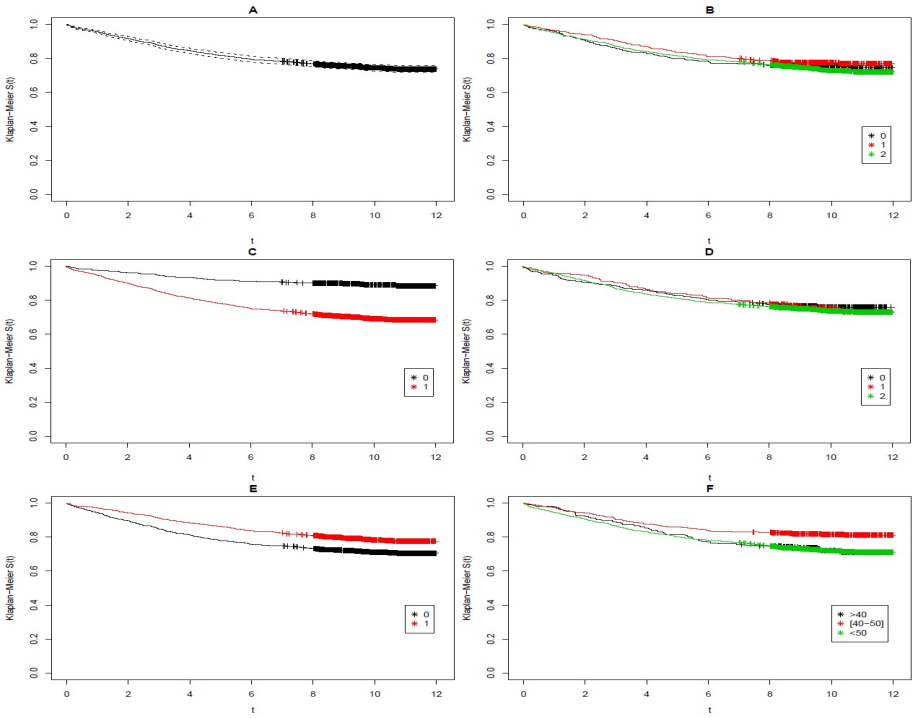

Figure 1 - Kaplan-Meier survival curves: (A) All patients, with 95% confidence intervals, (B) ICD (2: C509, 1: C504, 0: Others), (C) Race (1: White, 0: Others), (D) Morphology (2: Ductal Carcinoma, 1: Lobular Carcinoma, 0: Others), (E) Marital status (1: Married, 0: Others) and (F) Age group of women diagnosed with breast cancer in Parana State from 2004 to 2016.

3.2 Analysis with long-term survivors

The Weibull cure survival model (2) under a Bayesian perspective was employed here for the data analysis of breast cancer patients from Parana State, observed from 2004 to 2016. In order to find some risk factors that have influence in the death risk of breast cancer, we fitted the proposed models in (2) with several available covariates such as Age group (A), Morphology (M), Topography code (ICD), Marital status (MS), and Race (R). The corresponding Weibull regression with these covariates is here represented by the following structure for its scale parameter

$$\lambda \equiv \mathrm{e}^{\eta} = \exp\left[\beta_0 + \sum_{j=1}^{2}\left(\beta_j A_j + \beta_{j+2} M_j + \beta_{j+4} ICD_j\right) + \beta_7 MS + \beta_8 R\right], \quad (5)$$

where the corresponding dummy variables are defined by $A_1 = 1$ $(40 - 50$ years), $A_2 = 1$ $(> 50$ years), $M_1 = 1$ (lobular carcinoma), $M_2 = 1$ (ductal carcinoma), $ICD_1 = 1$ (C504), $ICD_2 = 1$ (C509), $MS = 1$ (married patient), and $R = 1$ (white race).

For defining the posterior distribution (4), the prior information considered was non-informative prior distributions since it was not possible to obtain prior information on the model parameters from past studies or from the opinion of experts on this topic. In this sense, the vague but proper prior distributions were standard Normal $\mathcal{N}(0, 1)$ for all regression parameters β_j, $j = 0, 1, \ldots, 8$ and Gamma $\mathcal{G}(1, 1)$ for the shape parameter γ. For the convenience of computational implementation only, some parameters were transformed, that is, $\theta_1 = \log \gamma$ and $\theta_2 = \log\left[\frac{p}{(1-p)}\right]$.

In order to compare the cure survival models, the DIC, CPO and WAIC measures were used. DIC, CPO and WAIC values of the models, with and without a cure fraction and selected covariates, are in Table 2. These values indicate that in fact some covariates are influential in the survival times of the patients with breast cancer, as well as that cure models are better fit than models without cure fraction. This will be discussed further ahead. Among these models, the selected model is the proposed Weibull survival model with long-term survivors and some covariates.

Table 2 - DIC, CPO and WAIC values of breast cancer survival models with and without long-term survivors and some covariates

Model		DIC	CPO	WAIC
No long-term survivors	No covariates	4,667.74	1,184.32	4,667.57
	With covariates	4,532.92	1,202.32	4,533.07
With long-term survivors	No covariates	4,646.37	1,187.34	4,646.70
	With covariates	4,521.26	1,214.38	4,522.51

Inferences on the parameters of both survival models with and without long-term survivors were also obtained through the package R-INLA (RUE *et al.*, 2009) using the families of Weibull and Weibull-cure distributions, respectively. Some inferential results of the selected model are in Table 3 such as posterior mean, standard deviation, and 95% HPD (High Posterior Density) credible intervals. It is noted that there is evidence of influence of Age group, Marital status and Race in the survival times of breast cancer patients (see further comments below).

Table 3 also shows that the proportion of long-term survivors (p) in the population of women with breast cancer was estimated at approximately 58%, based on the Weibull model with Age group, Race, Marital status, Topography code and Disease morphology. In the selected model, white race has a positive effect on the variable response i.e. white women with breast cancer have a higher risk of disease death than women of other races. Patients diagnosed with breast cancer between 40 and 50 years have a lower risk of death than women over 50 years, while married

Table 3 - Posterior quantities: mean, standard deviation (SD) and 95% HPD credible intervals (CI) for the selected cure model parameters

Parameter	Mean	SD	95% HPD IC
β_0 (intercept)	-8.525	0.405	(-9.322, -7.739)
β_1 ($40-50$ years)	-0.604	0.231	(-1.050, -0.145)
β_2 (>50 years)	-0.157	0.198	(-0.537, 0.239)
β_3 (lobular carcinoma)	-0.041	0.222	(-0.481, 0.389)
β_4 (ductal carcinoma)	0.081	0.142	(-0.196, 0.360))
β_5 (C504)	-0.131	0.202	(-0.528, 0.266)
β_6 (C509)	0.124	0.171	(-0.212, 0.460)
β_7 (married)	-0.722	0.130	(-0.977, -0.468)
β_5 (white)	1.650	0.167	(1.329, 1.983)
γ (shape parameter)	0.932	0.045	(0.844, 1.020)
p (cure fraction)	0.576	0.028	(0.519, 0.629)

patients have a lower risk of death than women who are not married. Another important factor is that patients with ICD C504 and C509 are not significant in the selected model, but patients with C504 have a negative 'residual' effect on the survival times, unlike the positive effect of patients with C509. This was confirmed in informal contact with a local oncologist who stated that ICD did not indicate a difference in the location of breast cancer.

Figure 2 presents the survival functions of breast cancer patients, estimated by Kaplan-Meier curve (1) and Weibull regression models with (2) and without (3) long-term survivors. Once the 'best' survival model (2) has long-term survivors, we concluded that the non-cure fraction survival model (3) is overestimating the breast cancer survival function, as well as the Kaplan-Meier curve. So, the cure fraction is below the level previously thought based on both the non-cure Weibull model and empirical survival function.

Figure 3 shows the residual plots of the selected model from a Bayesian perspective (see definitions in Wang *et al.*, 2018). Bayesian Cox-Snell, deviance and Martingale residual plots suggest that fitting the Weibull cure survival model is essentially good. Cox-Snell residuals are around the 45^o line, except from the middle to the tail, where the variability of the cumulative hazard function estimate is large. In addition, there is the presence of two groups of residual values in martingale and deviance residual plots. However, this question should be researched in the future because the definition of residual itself should vary between 'cured' and 'uncured' patients. Concerning these residual plots for all models in Table (2), they did not bring news beyond the distinction between these models depicted in the model comparison measures in this table. So, we chose to keep only the residuals plots of the selected model in current figure.

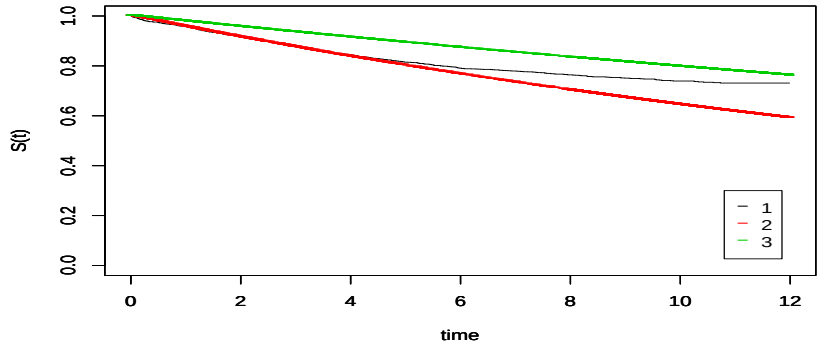

Figure 2 - Breast cancer survival functions, estimated by Kaplan-Meier curve (1) and Weibull models with (2) and without (3) long-term survivors.

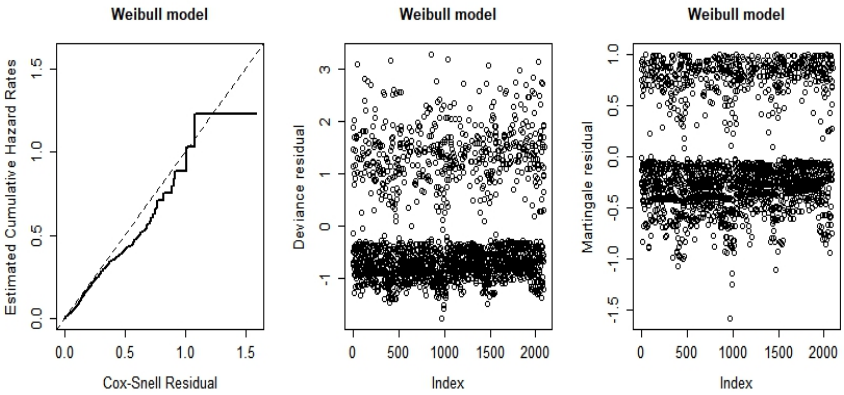

Figure 3 - Residual plots for diagnosing the selected Weibull cure model.

4 Conclusion

The proposed survival analysis for breast cancer data involving female patients of the Parana State, Brazil, from 2004 to 2016 was shown to be more appropriate because it includes a proportion of long-term survivors in the Weibull survival model in order to capture the presence of long-term survivors among the patients in the study population. This 'cure' fraction was estimated by approximately 58%. Based on the results obtained, the covariates Race, Age group and Marital status also had 'significant' effects on the risk of death from breast cancer.

Patients diagnosed with breast cancer between 40 and 50 years stand out from

other age groups in a positive way i.e. they have less death risk from breast cancer. Indeed a younger age at diagnosis is linked to increased mortality, according to McGuire *et al.* (2015) who also point out digital mammography is the superior method for detecting breast cancer, except in young patients. This is reflected in practice that method is the most commonly used diagnostic tool for older patients ($>$ 50 years old). In addition, Chen *et al.* (2016) indicate that younger breast cancer patients exhibit more aggressive disease than older patients, whereas middle-aged patients exhibit better survival than young and elderly patients.

Married patients have a positive effect in the breast cancer survival curve. Although there are no many studies about that issue, Hinyard *et al.* (2017) said marital status is strongly associated with improved health and longevity, and unmarried women were 1.2 times more likely to be diagnosed at a larger stage than married women. Perhaps this is linked to endocrine factors i.e. married women usually breastfeed their children. Based on a meta-analysis of observational studies, Li *et al.* (2020) corroborate this result by stating that insufficient exploration of confounding effects or inadequate ascertainment of marital status may limit the quality of that evidence.

White women with breast cancer have a higher risk of death than women of other races. This is the least studied issue so far, with more information on the population of the USA. Curtis *et al.* (2008) revealed that African American women had worse survival than white women, although controlling for predictor variables reduced this difference among all stage breast cancer. However, they also proposed further investigate the role of biology, demographics, and disparities in quality of care. Chatterjee *et al.* (2013) stated that black women are more often diagnosed with advanced-stage disease than are White women, concluding that in the mammography era, racial disparities is stage at breast cancer diagnosis in the USA.

We share some of the concerns and conclusions above, especially wishing there is more research on breast cancer risk factors in future. Nevertheless, in studies based on population oncological records in Brazil and not only, the results of cancer data analysis should be looked at with some care since there is a need to check the quality of collected data and the excess of missing observation, being impossible to control this during the data analysis process.

Although MCMC methods are widely used in Bayesian analysis, it was also more convenient to employ INLA approach for obtaining inferential results, since INLA has already implemented the Weibull-cure distribution and a faster execution of computational codes (here executed in R-INLA). On the other hand, INLA has not yet implemented important issues for a thorough analysis of data with cure fraction, for instance, including covariates into the fraction of long-term survivors, predicting new survival times based on the current distributions, and varying the values fixed for the parameters of the prior distributions. These are certainly topics for extensions of this work in future.

Recent studies involving more innovative methodology in the processing of survival data by using cure models are stimuli for future work of cure survival

models for the breast cancer data analysis. For example, Fernandes *et al.* (2018) came up with a survival model with a cure rate in a scenario with M risk factors, considering the discrete Lindley distribution for M and the Weibull distribution for the activation time of each factor. Wei and Wu (2019) proposed a cure model with proportional risks by parts to incorporate the effect of delayed treatment and cure fraction in clinical trials of cancer immunotherapy. Seppä *et al.* (2019) used overmortality models with random effects to estimate the variation in relative survival or net survival of patients with cancer. This study evaluated the performance of INLA in monitoring regional variation for cancer registration data. Lastly, caution on the general use of these results is nevertheless advised, since we have analyzed population-based breast cancer data without proper monitoring by a health professional.

Acknowledgments

The authors thank Dr. António E. Pinto (Instituto Português de Oncologia - IPO Lisboa) for his valuable suggestions, as well as the Editor-in-Chief and reviewers for the valuable comments. G.L. Silva was partially funded by Fundação para a Ciência e a Tecnologia (FCT-Portugal) project UIDB/00006/2020. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

MORAES, T. E. N. T.; PREVIDELLI, I. T.; SILVA, G. L. Uma análise de regressão Weibull bayesiana de dados de câncer de mama com sobreviventes a longo prazo no Estado do Paraná, Brasil. *Rev. Bras. Biom.,* Lavras, v.39, n.2, p.293-310, 2021.

- *RESUMO: O câncer de mama é uma das doenças mais comuns entre as mulheres em todo o mundo, com cerca de* 25% *de novos casos a cada ano. No Brasil, eram esperados 59,7 mil novos casos de câncer de mama em 2019, segundo o Instituto Nacional do Câncer (INCA). A análise de sobrevivência tem sido uma ferramenta útil para identificar os fatores de risco e prognóstico para pacientes com câncer. Este trabalho tem como objetivo caracterizar o valor prognóstico de variáveis demográficas, clínicas e patológicas em relação ao tempo de sobrevivência de 2092 pacientes com diagnóstico de câncer de mama no Estado do Paraná, Brasil, de 2004 a 2016. Nesse sentido, propomos uma análise bayesiana de dados de sobrevdvência com sobreviventes a longo prazo usando modelos de regressão Weibull por meio de aproximações de Laplace encaixadas e integradas (INLA). Os resultados apontam para uma proporção de sobreviventes a longo prazo em torno de* $57,6\%$ *na população em estudo. Em relação aos potenciais fatores de risco, concluímos nomeadamente que a faixa etária de 40-50 anos tem sobrevivência superior aos grupos etários mais jovens e mais velhas, as mulheres brancas têm maior risco de câncer de mama do que outras raças e o estado civil diminui esse risco. No entanto, recomenda-se cautela no uso geral desses resultados, uma vez que analisamos dados de câncer de mama de base populacional sem o devido monitoramento de um profissional de saúde.*

- *PALAVRAS-CHAVE: Regressão Weibull; Fração de cura; INLA; Estatística bayesianas; Análise de sobrevivência.*

References

ACHCAR, J. A.; COELHO-BARROS, E. A.; MAZUCHELI, J. Cure fraction models using mixture and non-mixture models. *Tatra Mountains Mathematical Publications*, v.51, n.1, p.1-9, 2012.

AMARAL-TURKMAN, M. A.; PAULINO, C. D.; MÜLLER, P. *Computational Bayesian statistics: An introduction*. Cambridge University Press, 2019, 243p.

AMICO, M.; VAN KEILEGOM, I. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, v.5, p.311-342, 2018.

APARÍCIO, M. M. E. F. *Análise de sobrevivência do cancro da mama masculina*. 2013. 120p. Dissertation (Master Degree in Biomedical Engineering) - Instituto Superior Técnico, Universidade de Lisboa, Portugal, 2013.

ARAÚJO, I. M. A.; FERNANDES, A. F. C. O significado do diagnóstico do câncer da mama para a mulher. *Escola Anna Nery - Revista de Enfermagem*, v.12, n.4, p.664-671, 2008.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v.47, n.259, p.501-515, 1952.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society B*, v.11, n.1, p.15-53, 1949.

CARROLL, R.; LAWSON, A. B.; FAES, C.; KIRBY, R.S.; AREGAY, M.; WATJOU, K. Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spatial and Spatio-temporal Epidemiology*, v.14-15, p.45-54, 2015.

CHATTERJEE, N. A.; HE, Y.; KEATING, N. L. Racial differences in breast cancer stage at diagnosis in the mammography era. *American Journal of Public Health*, v.103, n.1, p.170-176, 2013.

CHEN, H-l.; ZHOU, M-q.; Tian, W.; MENG, K-x.; HE H-f. Effect of age on breast cancer patient prognoses: A population-based study using the SEER 18 Database. *PLoS ONE*, v.11, n.10, e0165409, 2016.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blücher, 2006, 392p.

COX, D. R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, v.34, p.187-220, 1972.

CUCCHETTI, A.; FERRERO, A.; CESCON, M.; DONADON, M.; RUSSOLILLO, N.; ERCOLANI, G.; STACCHINI, G.; MAZZOTTI, F.; TORZILLI, G.; PINNA, A. D. Cure model survival analysis after hepatic resection for colorectal liver metastases. *Annals of Surgical Oncology*, v.22, n.6, p.1908-1914, 2015.

CURTIS, E.; QUALE, C.; HAGGSTROM, D.; SMITH-BINDMAN, R. Racial and ethnic differences in breast cancer survival. How much is explained by screening, tumor severity, biology, treatment, comorbidities, and demographics? *Cancer*, v.112, n.1, p.171-180, 2008.

DE SMEDT, T.; SIMONS, K.; NIEUWENHUYSE, A. V.; MOLENBERGHS, G. Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. *Archives of Public Health*, v.73, S.1, O.2, 2015.

FERNANDES, P. G.; SUZUKI, A. K.; SARAIVA, E. F. O modelo Lindley-Weibull com proporção de cura: uma abordagem bayesiana. *Revista Brasileira de Biometria*, v.36, n.4, p.998-1022, 2018.

HINYARD, L.; WIRTH, L. S.; CLANCY, J. M.; SCHWARTZ, T. The effect of marital status on breast cancer-related outcomes in women under 65: A SEER database analysis. *The Breast*, v.32, p.13-17, 2017.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. *Bayesian survival analysis*. Wiley, 2001.

INCA - Instituto Nacional de Câncer José Alencar Gomes da Silva, Ministério da Saúde, Brasil. *A situação do câncer de mama no Brasil: Síntese de dados dos sistemas de informação*, 2019.

LAMBERT, P. C.; THOMPSON, J. R.; WESTON, C. L.; DICKMAN, P. W. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, v.8, n.3, p.576-594, 2006.

LAMBERT, P. C. Modeling of the cure fraction in survival studies. *The Stata Journal*, v.7, n.3, p.351-375, 2007.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. 2.ed. New York: Wiley, 2003.

LEE, E. T.; WANG, *Statistical methods for survival data analysis*. 3.ed. New Jersey: Wiley, 2003.

LI, M.; HAN, M.; CHEN, Z.; TANG, Y.; MA, J.; ZHANG, Z.; LIU, Z.; ZHANG, N.; XI, C.; LIU, J.; TIAN, D.; WANG, X.; HUANG, X.; CHEN, J.; WANG, W.; ZHAI, S. Does marital status correlate with the female breast cancer risk? A systematic review and meta-analysis of observational studies. *PLoS ONE*, v.15, n.3, e0229899, 2020.

MALLER, R. A.; ZHOU, X. *Survival analysis with long-term survivors*. New York: Wiley, 1996.

MARTINEZ, E. Z.; ACHCAR, J. A.; JÁCOME, A. A. A.; SANTOS, J. S. Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. *Computer Methods and Programs in Biomedicine*, v.112, n.3, p.343-355, 2013.

MCGUIRE, A.; BROWN, J. A. L.; MALONE, C.; MCLAUGHLIN, R.; KERIN, M.J. Effects of age on the detection and management of breast cancer. *Cancers*, v.7, p.908-929, 2015.

PAULINO, C. D.; AMARAL-TURKMAN, M. A.; MURTEIRA, B.; SILVA, G. L.*Estatística bayesiana*. 2.ed. Lisbon: Fundação Calouste Gulbenkian, 2018, 601p.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: `https://www.R-project.org/`.

RAMIRES, T. G.; CORDEIRO, G. M.; KATTAN, M. W.; HENS, N.; ANDORTEGA, E. M. Predicting the cure rate of breast cancer using a new regression model with four regression structures. *Statistical Methods in Medical Research*, v.27, n.11, p.3207-3223, 2018.

RODRIGUES, J. D.; CRUZ, M. S.; PAIXÃO, A. N. Uma análise da prevenção do câncer de mama no Brasil. *Ciência & Saúde Coletiva*, v.20, p.3163-3176, 2015.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, v.71, n.2, p.319-392, 2009.

SEPPÄ, K.; RUE, H.; HAKULINEN, T.; LÄÄRÄ, E., SILLANPÄ, M. J.; PITKÄNIEMI, J. Estimating multilevel regional variation in excess mortality of cancer patients using integrated nested Laplace approximation. *Statistics in Medicine*, v.38, n.4, p.778-791, 2019.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; VAN DER LINDE, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, v.64, n.4, p.583-639, 2002.

SPOSTO, R. Cure model analysis in cancer: An application to data from the children's cancer group. *Statistics in Medicine*, v.21, n.2, p.293-312, 2002.

SUGUIURA, T. P. S. *Modelos multiníveis: gaussiano e multinomial*. 2017. 128p. Dissertation (Master Degree in Biostatistics) - Universidade Estadual de Maringá, Brazil, 2017.

WANG, X.; RYAN, Y. Y.; FARAWAY, J. J. *Bayesian regression modeling with INLA*. Boca Raton: Chapman and Hall/CRC Press, 2018.

WATANABE, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, v.11, p.3571-3594, 2010.

WEI, J.; WU, J. Cancer immunotherapy trial design with cure rate and delayed treatment effect. *Statistics in Medicine*, v.39, n.6, p.698-708, 2020.

YAKOVLEV, A. Y.; TSODIKOV, A. D. *Stochastic models of tumor latency and their biostatistical applications*. Singapore: World Scientific, 1996, 288p.

Received on 24.03.2020.

Approved after revised on 07.12.2020.

Appendix

A R-INLA code for the selected model in Table 2

```
data1 = read.csv('cancer4years.csv', sep = ',',  header = T,
                 na.strings = "NA")
require(INLA)

formula1 = inla.surv(time, censoring) ~ age.group + topography +
           morphology + marital.status

model1 = inla(formula1, family="weibullcure", data=dados1,
        control.compute=list(dic=TRUE, waic=TRUE, cpo=TRUE))

summary(model1)
```