Mathematical Biosciences and Engineering

https://www.aimspress.com/journal/mbe

MBE, 23(7): 1869–1885.
DOI: 10.3934/mbe.2026068
Received: 04 March 2026
Revised: 12 May 2026
Accepted: 22 May 2026
Published: 30 June 2026

Research article

Survival analysis of women diagnosed with breast cancer in Angola

Jaime A. Jerónimo[1,2], **Dulce Gomes**[2,3] **and Patrícia A. Filipe**[2,4,*]

[1] Universidade Católica de Angola and Instituto Superior de Administração e Finanças (ISAF), Luanda, Angola

[2] Centro de Investigação em Matemática e Aplicações (CIMA), Instituto de Investigação e Formação Avançada, Universidade de Évora, Évora, Portugal

[3] Departamento de Matemática, Escola de Ciência e Tecnologia, Universidade de Évora, Évora, Portugal

[4] Departamento de Métodos Quantitativos para Gestão e Economia, Iscte Business School and Business Research Unit (BRU-IUL), Iscte-Instituto Universitário de Lisboa, Lisboa, Portugal

* **Correspondence:** Email: patricia.filipe@iscte-iul.pt; Tel: +351217903000.

Abstract: Breast cancer is the most common cancer among women and represents 24.5% of all cancer cases worldwide. As there are few studies on the subject in Angola, it is particularly interesting to characterize the survival of women diagnosed with breast cancer in that country. The collected data set contains all diagnosed patients followed up at the Angolan Cancer Control Institute (IACC) between 2013 and 2022. However, due to the impact of COVID-19, the data set under analysis consists of the prepandemic period (2013 to 2019). The aim is to study the population diagnosed with cancer in the country, not only descriptively but also to check which factors influence the survival time of patients from diagnosis to death from breast cancer. To this end, various methods are applied to see which best describes the time until the event under study occurs. Namely, this study examines survival analysis models, beginning with the well-known Cox proportional hazards model. It then transitions to parametric models based on exponential and Weibull distributions and finally explores generalizations of these models that offer greater flexibility. To achieve this flexibility in the survival distribution, we adopt the proportional hazards framework proposed by Younes and Lachin but applying a new approach by Royston and Parmar. The results show that the median survival time from breast cancer in the observed women was 560 days. The variables that were revealed to be significant risk factors were age, stage of the disease, and body mass index. According to Akaike's criterion, the flexible proportional hazards model with one knot proved to be the most appropriate model for these data.

Keywords: breast cancer; parametric models; semiparametric models; Royston–Parmar models; survival analysis

1. Introduction

The incidence of cancer in Angola is one of our main social concerns, and knowledge of the epidemiological profile of cancer is fundamental to planning national policies related to this disease. In Angola, the available data indicate that around 15,000 cancer cases are registered every year, with an estimated 10,000 deaths. The most common types of cancer among the female population include breast (20.8%) and cervical (20.3%) [1].

Recent data show that the cancer situation in Africa is disheartening. In 2022 alone, there were around 882,882 new cases of cancer in the African region of the World Health Organization (WHO), with around 573,653 deaths. If urgent action is not taken, cancer mortality in the region is expected to reach around one million deaths a year by 2030 [2].

According to WHO experts, breast cancer has overtaken lung cancer as the most frequently diagnosed type of cancer in the world. This change was reported in statistics released by the International Agency for Research on Cancer (IARC) in December 2020. More recent data reveal that lung cancer is the most common worldwide but followed by breast cancer [3]. Breast cancer is a public health priority and is responsible for one in six deaths among women worldwide. Particularly in Angola, the most frequently diagnosed cancer for both sexes (11.8%) is breast cancer, and lung disease is not even in the top five most diagnosed cancer types in the country [1].

In addition to the direct effects of cancer on the health of the individual, the disease has a major impact on society in general. More specifically, because it mainly affects the working population, it jeopardizes the development of the community, as this is the sector of the population that ensures both the production of wealth and demographic balance.

For developing countries such as Angola, this is a major problem due to incipient health policies; almost nonexistent prevention measures; and the scarcity, or even nonexistence, of means of treatment, the costs of which are burdensome and restrict better results in terms of control.

We intend to promote the relevance of a study on the survival analysis of women diagnosed with breast cancer in Angola and to contribute to a better understanding of this reality that is severely affecting Angolan families and the community. The aim is to study, analyze, and characterize the women fighting breast cancer in the country and identify possible risk and/or protective factors, examining which aspects are affecting their survival time. For this purpose, a sample consisting of all women diagnosed with breast cancer and followed at the IACC between the years 2013 and 2022, with data distributed across both qualitative and quantitative variables, was collected. An impact of the COVID-19 pandemic was observed, where, in 2020, the number of patients who sought health services for breast cancer diagnosis decreased by 173 (-57%) compared to the corresponding data for 2019 (prepandemic period). The results indicated a significant reduction in the number of breast cancer diagnostic procedures performed on patients who should have been seen at the IACC. To avoid the bias caused by the pandemic period, only information up to 2019 was used for the this study. Section 2 presents the techniques considered most appropriate for this research, namely survival analysis models. The well-known Cox proportional hazards semiparametric model is presented, followed by parametric models based on exponential and Weibull distributions for the survival time. The exponential model, which represents a particular case of the Weibull distribution, was included as a reference model to facilitate comparison across parametric and semiparametric approaches, followed by the flexible proportional hazards model introduced by Younes and Lachin but applying a new approach by Royston

and Parmar. Section 3 presents the application and the comparison of the survival analysis methods that identify which variables are statistically significant in explaining the survival time (in days) of women with breast cancer in Angola in the period 2013–2019 from the date of disease diagnosis until death. Section 4 contains the main conclusions.

2. Survival analysis models

Survival analysis techniques are well-established in a wide range of fields. In particular, to our knowledge, there are few studies in the literature that have applied flexible Royston–Parmar models to breast cancer survival. In the specific case of our application to breast cancer in Angola, there are no publications in the literature, either for this particular type of model or even for the most commonly used survival analysis models.

In this section, we will begin by introducing some general concepts and describing the most commonly used models, such as the Cox semiparametric model and other parametric models, in order to provide a framework for this type of analysis and prepare the reader for the description of the Royston–Parmar flexible models.

2.1. General concepts

In survival analysis, time is a non-negative continuous random variable T with a certain distribution function $F(t)$ and the respective probability density function $f(t)$. The survival function $S(t)$ represents the probability that a population will survive at least until time t and is defined as

$$S(t) = P(T > t). \tag{2.1}$$

The instantaneous hazard function $h(t)$, which expresses the instantaneous mortality rate of an individual that survives until time t, is obtained by dividing the density function $f(t)$ and the survival function $S(t)$ according to the equation:

$$h(t) = \frac{f(t)}{S(t)}, \tag{2.2}$$

and the cumulative hazard function is given by

$$H(t) = \int_0^t h(v)dv, \tag{2.3}$$

which is therefore a measure of the risk that the event occurs at a given specific time.

Consider that in addition to the survival times of the individuals, the covariates associated with these individuals will also be considered. The aim, as with other regression models, is to identify which covariates are related to the event under study and how these affect the survival function. Therefore, let $\mathbf{X}$ be the vector of covariates and

$$S(t|\mathbf{x}) = P(T > t|\mathbf{X} = \mathbf{x}) \tag{2.4}$$

the survival function, conditioned on the covariates vector. This conditional survival function relates to the cumulative hazard function, $H(t|\mathbf{x})$, through the following identity:

$$H(t|\mathbf{x}) = -\ln(S(t|\mathbf{x})). \tag{2.5}$$

2.2. Semi-parametric models

The Cox model, also known as the Cox proportional hazards model, is currently the most widely used procedure for modelling the relationship between a set of covariates and survival time T but without requiring any probability distribution to be assumed for T.

One of its assumptions is that the hazards are proportional. In other words, the ratio of hazards between groups is constant over time. In the Cox model, the hazard function at time t is defined by

$$h(t|\mathbf{x}) = h_0(t)\exp(\beta^\intercal \mathbf{x}), \tag{2.6}$$

where $h_0(t)$ is the baseline hazard function, that is, the risk in the absence of covariates, $\mathbf{x} = (x_1, x_2, ..., x_p)$ is the vector of p covariates, and $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is the vector of parameters.

Applying the logarithm to the hazard function results in

$$\ln(h(t|\mathbf{x})) = \ln(h_0(t)) + \beta_1 x_1 + \cdots + \beta_p x_p, \tag{2.7}$$

that is, a log-linear model is obtained, which makes it much easier to estimate the hazard function.

The survival function is therefore written as

$$S(t|\mathbf{x}) = S_0(t)^{\exp(\beta^\intercal \mathbf{x})}, \tag{2.8}$$

where $S_0(t)$ represents the baseline survival function.

The ratio between the risk of the event occurrence between two patients i and j, with covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ e $\mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{jp})$, is therefore given by

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} = \frac{\exp(\beta^\intercal \mathbf{x}_i)}{\exp(\beta^\intercal \mathbf{x}_j)} = \exp(\beta^\intercal(\mathbf{x}_i - \mathbf{x}_j)), \tag{2.9}$$

independent of t.

Assuming hazards proportionality, the Cox model can estimate the covariates effects without considering any specific distribution for the survival time T. Model parameters β are estimated using the partial likelihood method proposed by Cox [4]. Under the assumption of no tied event times, the estimator $\hat{\beta}$ is obtained by maximizing the partial likelihood function

$$L(\beta) = \prod_{i \in D} \frac{\exp(\beta^\intercal \mathbf{x}_i)}{\sum_{j \in R(t_i)} \exp(\beta^\intercal \mathbf{x}_j)}, \tag{2.10}$$

where D represents the set of indices of patients who experience the event, t_i denotes the event time for patient i, and $R(t_i)$ is the risk set at time t_i (all patients who have not yet experienced the event or been censored just prior to t_i).

2.3. Parametric models

Parametric models are models that assume a specific continuous probability distribution for the survival time T. Various distributions can be used in this type of model. We present here two forms, one being the simplest parametric model, when the survival time T has an exponential distribution, and the other being one of the most popular models, where T has a Weibull distribution, in order to make the brief summary of flexible parametric models presented here understandable.

The exponential distribution is appropriate for the case where the hazard is constant over time. Assuming that T follows an exponential distribution where the parameter λ depends on the vector of covariates $\mathbf{x}$, the survival and hazard functions are given, respectively, by

$$S(t|\mathbf{x}) = P(T > t|\mathbf{x}) = \exp(-\lambda(\mathbf{x})t), \text{ where } \quad \lambda(\mathbf{x}) = \exp(\beta_1 x_1 + \cdots + \beta_p x_p) \tag{2.11}$$

and

$$h(t|x) = \frac{f(t|\mathbf{x})}{S(t|\mathbf{x})} = \lambda(\mathbf{x}). \tag{2.12}$$

Thus, in this case, applying the logarithm to the hazard function, a linear function of the covariates is obtained:

$$\ln(h(t|\mathbf{x})) = \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2.13}$$

The exponential model is the simplest, but it has some practical limitations. For example, the hazard function is constant over time. One of the most widely used distributions that overcomes this limitation is the Weibull distribution.

The Weibull distribution has two positive parameters, λ and γ, which are, respectively, the scale and shape parameters, and we can also assume a vector of covariates $\mathbf{x}$. The hazard function is given by

$$h(t|\mathbf{x}) = \lambda^{-\gamma}\gamma t^{\gamma-1}\exp(\beta^\intercal \mathbf{x}), \quad t \geq 0. \tag{2.14}$$

If $\gamma > 1$, this means that the hazard is increasing over time, and if $(0 < \gamma < 1)$, it means that the hazard is decreasing over time. The particular case in which $\gamma = 1$ corresponds to the exponential distribution.

As mentioned in (2.6), in the context of proportional hazards models, the baseline hazard function takes the form

$$h_0(t) = \lambda^{-\gamma}\gamma t^{\gamma-1}, \quad t \geq 0, \tag{2.15}$$

and thus,

$$\ln(H(t|\mathbf{x})) = \ln\left[\left(\frac{t}{\lambda}\right)^{\gamma}\right] + \beta^\intercal \mathbf{x}. \tag{2.16}$$

For an introduction to and better understanding of Roystan and Parmar's flexible models [5], it is defined as $t' = \ln(t)$, $\alpha_0 = -\gamma\ln(\lambda)$, and $\alpha_1 = \gamma$, leading to

$$\ln(H(t|\mathbf{x})) = \gamma\ln(t) - \gamma\ln(\lambda) + \beta^\intercal \mathbf{x} = \alpha_0 + \alpha_1 t' + \beta^\intercal \mathbf{x}, \tag{2.17}$$

which is a linear function of t' through a function $s(t', \alpha)$ with $\alpha = (\alpha_0, \alpha_1)$. The same linear relation is observed for the cumulative odds function, using a log-logistic distribution.

2.4. Royston–Parmar flexible model

Parametric models provide not only an estimate of the baseline hazard function but also estimates of the coefficients associated with the covariates. However, "the estimation of the baseline risk has long been neglected. However, the effect of the hazard ratio can only be fully understood if the baseline hazard is known as well" (chapter 1, page 9, [6]).

Due to technical difficulties in the estimation of the baseline hazard function, the semiparametric proportional hazard model (Cox model) appears as the most applied model.

The flexible models proposed by Younes and Lachin [7] are, as the name suggests, a flexible generalization transformation of the survival function through a link function, which the authors called $g(.)$, known, bijective, and strictly monotone, such that $g : (0,1) \to (-\infty, \infty)$. These authors discuss that "in practice, the analyst needs flexibility in modeling not only the background [baseline] hazard or survival but also the effect of covariates on survival. Models such as the proportional hazards model, the proportional odds model, or the accelerated failure time model make specific assumptions about the nature of this effect, specifically that the covariates act multiplicatively on the background [baseline] hazard, the background [baseline] odds of survival, or the expected survival time, respectively" (Section 1, page 1199, [7]).

The flexible models consist of modelling a transformation of the survival function $S(t)$ as a link function $g(.)$ according to the equation bellow:

$$g[S(t|\mathbf{x})] = g[S_0(t)] + \beta^\intercal \mathbf{x}, \tag{2.18}$$

where $S_0(t)$ is the baseline survival function, and β is the vector of parameters to be estimated for the covariates $\mathbf{x}$.

The authors of [5] and [7] used two link functions of the parametric family proposed by Aranda-Ordaz in 1981 [8]:

$$g_\theta(\mathbf{x}) = \ln \frac{\mathbf{x}^{-\theta} - 1}{\theta}, \tag{2.19}$$

so that $\theta = 1$ corresponds to the proportional odds model and $\theta \to 0$ to the proportional hazards model.

Let us consider the relation between the survival function and the cumulative survival function

$$g[S(t|\mathbf{x})] = \ln(-\ln(S(t|\mathbf{x})) = \ln(H(t|\mathbf{x})) = \ln(H_0(t)) + \beta^\intercal \mathbf{x}. \tag{2.20}$$

Using the notation from the previous section,

$$g[S(t|\mathbf{x})] = \ln(H_0(t)) + \beta^\intercal \mathbf{x} = s(t', \alpha) + \beta^\intercal \mathbf{x}. \tag{2.21}$$

Younes and Lachin's approach used basis splines (or B-splines) to estimate the baseline hazard function (B-splines use the observed event times, divided into intervals by a set of internal knots, to construct a smooth, piecewise polynomial function. Each basis function acts locally over a specific interval of time, allowing the model to approximate the baseline hazard function with high flexibility) and then obtain the baseline survival function by integration. The new approach by Royston and Parmar, however, is based on the "logarithm of the baseline cumulative hazard function as a 'natural' cubic spline function of log time. Therefore, the general function [which we designate as] $s(t', \alpha)$ is to be approximated by a spline" [5].

Although the approach by Younes and Lachin is highly effective at capturing complex fluctuations in the baseline hazard function, the approach proposed by Royston and Parmar offers more constrained behavior at the boundaries of the follow-up period.

Further details, for example, on point or interval estimation and model selection can be found in [5].

3. Results

In this section, we present the results of the analysis of the survival time (in days) of 1306 women diagnosed with breast cancer between 2013 and 2019 in Angola, from the date of diagnosis of the disease until the occurrence of death or end of study. As covariates, we have age, age at menarche (onset of menstruation), age at menopause, stage of the disease (the classification recommended by the American Joint Committee on Cancer: grades ranging from 0 to IV), type of treatment submitted to patients (surgery/chemotherapy, palliative/hormone therapy and radiotherapy), job, marital status, risk factors (such as being a regular consumer of alcohol, a smoker, or being HIV positive), types of contraceptives used, and body mass index (BMI).

Initially, we used the WHO classification of women's BMI; however, due to the low percentage in the different obesity classes, we grouped BMI into just 3 categories: underweight, normal weight, and obese. The age variable was categorized into four groups: ≤ 39, $40-59$, $60-79$, and ≥ 80, according to typical age groups considered in other studies for breast cancer survival analysis (for instance, [9]). As for the stage variable, it was grouped into only three categories: up to grade II, grade III, and grade IV, also due to the fact that there were few patients diagnosed at stages 0, I, and II.

All the analysis were performed using `R` (version 4.3.2) [10], via the `RStudio` [11], and a significance level of 5% was considered.

Among the 1306 women studied, 476 (36.45%) died before the end of the follow-up period. The median survival time from breast cancer in the observed women was 560 days. 44.87% of the women diagnosed with breast cancer were in stage III of the disease at the time of diagnosis, and 11.18% were in stage IV (the most advanced cases according to the tumor–node–metastasis (TNM) classification of the American Joint Committee on Cancer).

The survival function estimation was done using the Kaplan–Meier estimator. The Kaplan–Meier estimator was used to estimate the cumulative probability of the event, accounting for right-censored observations [12]. Differences between the survival distributions of the different categories were assessed using the log-rank test, revealing a statistically significant or a nonsignificant difference between groups [13].

According to Figure 1, it can be seen that the estimated probability of a patient remaining alive until a given moment slowly decreases as the days since the cancer diagnosis increase, with the probability of survival establishing at around 15% at values above 2400 days. The corresponding 1°quartile value is 215 days, which means that at 215 days, the probability of survival is 75%; that is, at 215 days, the event of interest has already occurred in 25% of patients. The median and 3°quartile were 560 and 1688 days, respectively.

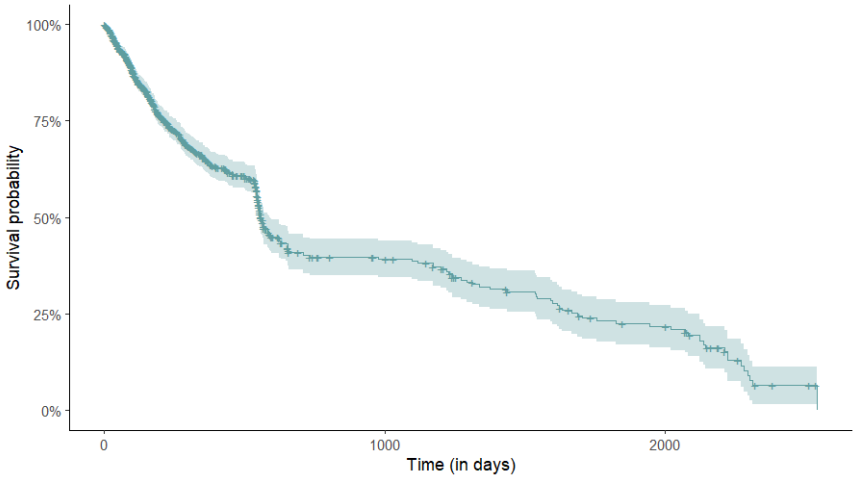

Figure 1. Kaplan–Meier curve of time from diagnosis to breast cancer death/end of study.

Kaplan–Meier estimates were obtained, and the log-rank test was performed on the categorical variables to test significant differences between categories. The results are shown in Figure 2 for the four variables that presented significant differences between groups.

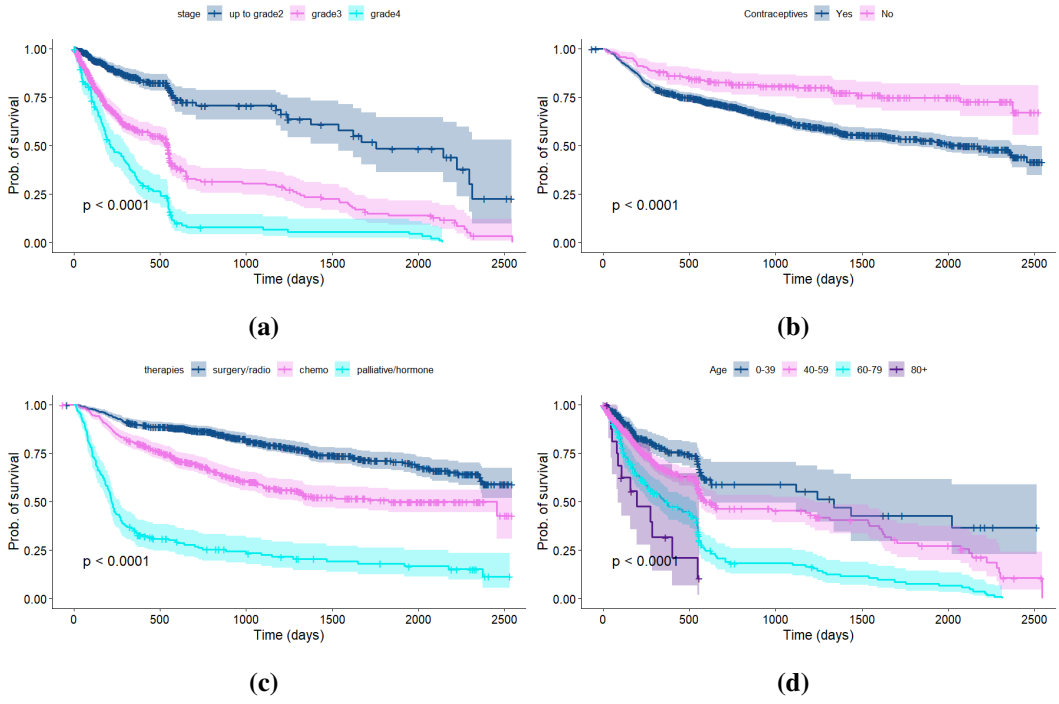

Figure 2. Kaplan–Meier curves and p-value for the log-rank test for the following variables: (**a**) stage, (**b**) contraceptives, (**c**) therapies, and (**d**) age.

Looking at Figure 2(a), according to the results of the log-rank test, the corresponding survival

curves are significantly different (p-value < 0.001). It has been found that cases diagnosed at stage IV and III have a lower chance of survival compared to a tumor at stage up to grade II for these patients.

Patients who used contraceptive methods, as shown in Figure 2(b), were more likely to die from breast cancer compared to patients who did not use such methods or did not respond. These findings should be interpreted through a lens of skepticism regarding data completeness. The fact that a staggering majority reported nonuse or declined to respond suggests that contraceptive use remains shrouded in social stigma or restricted by limited accessibility, effectively rendering the reported population that use contraceptives 15% (12.5% ages < 39) a possible marginal representation of the Angolan population's realities.

Patients undergoing chemotherapy have a higher risk of dying from breast cancer compared to patients undergoing surgery/radiotherapy and palliative/hormonotherapy, according to Figure 2(c).

Finally, according to Figure 2(d), the probability of surviving breast cancer decreases as the age of the patients increases at the time of diagnosis of the disease: that is, older patients who were between 60–79 and over 80 years old at the time of diagnosis have a lower probability of surviving breast cancer when compared to the other ages.

3.1. Cox semiparametric model

The Cox semiparametric model was applied, and the significant variables were identified. With this purpose, we have started by including all the categorical and continuous variables (age was included as a continuous variable) under study in the model. Subsequently, in each of the following steps, the variables were eliminated according to their p-value. In other words, the procedure was to eliminate, one by one, the nonsignificant variables with the highest p-value and calculate the effect on the model using the likelihood ratio test to assess whether the model without the variable is as explanatory as the full model. This procedure is carried out until all the variables in the model are significant. At this stage, a significance level α up to 10% can be considered, as advised by some authors such as Hosmer and Lemeshow [14].

Once the process of ascertaining the significance of the variables for survival time has been completed, the Cox model includes the following variables: age at diagnosis, disease stage, and body mass index. These are the variables responsible for the main effects on the time from diagnosis of the disease to the occurrence of the "death" event. Thus, we obtain a final model that contains only the variables stage, age, and BMI.

Table 1 and Figure 3 summarize the results of the Cox proportional hazards model, specifically showing it shows the estimated hazard ratio for the variables that proved to be significant, the respective 95% confidence intervals, and the p-values corresponding to the Wald test.

Table 1. Cox proportional hazards model: estimated coefficients, estimated hazard ratios, respective 95% confidence intervals (CI95%), and Wald test p-value.

Covariate	$\widehat{\beta}$	$\widehat{HR}$	*CI*95%	p-value
Age	0.028	1.029	$]1.022; 1.036[$	< 0.001
Stage (grade III)	1.054	2.871	$]2.239; 3.680[$	< 0.001
Stage (grade IV)	1.737	5.684	$]4.297; 7.518[$	< 0.001
BMI (overweight)	0.183	1.202	$]1.001; 1.444[$	0.0492

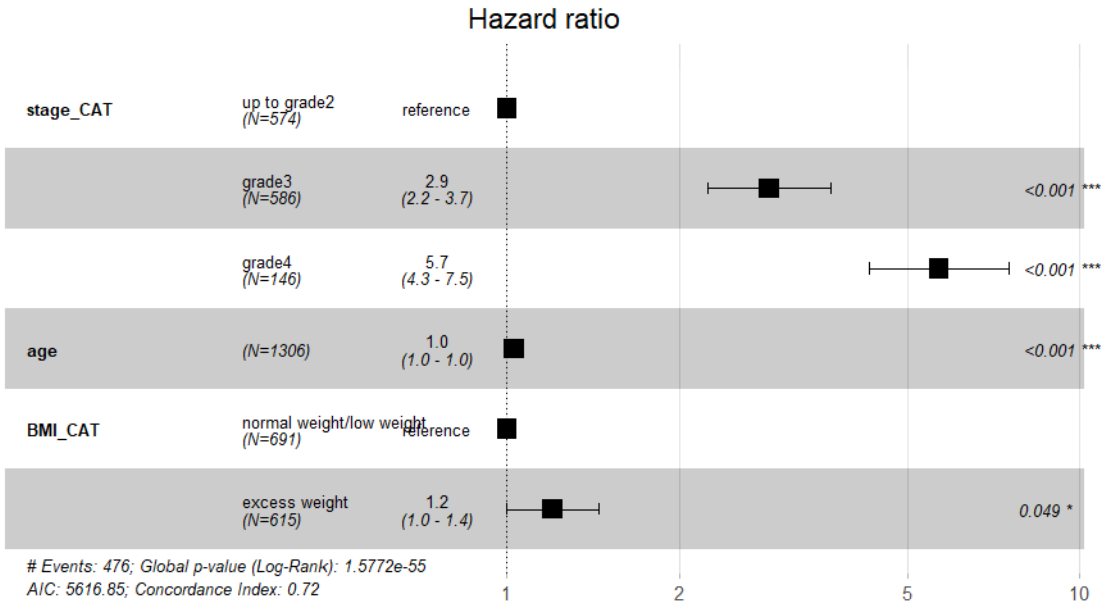

Figure 3. Summary of the results of the adjustment of the Cox proportional hazards model.

The time that elapses between the diagnosis of breast cancer and the moment when patients die is mainly influenced by the age at which they are diagnosed, the stage of the disease at diagnosis, and their body mass index. From the interpretation of the hazard ratio calculated for a significance level of 5% for each of the covariates that make up the explanatory model of survival time under study, considering the last category as a reference for categorical variables, it can be said that everything else remains constant:

(1) For every extra year a patient is diagnosed with breast cancer, her risk of death increases by between 2% and 3.6%, with 95% confidence;
(2) In the case of a patient who at the time of diagnosis was in stage III, their risk of death is almost three times higher than that of an individual who at the time of diagnosis was in stage II;
(3) On the other hand, a patient who at the time of diagnosis was in stage IV had a risk of death approximately 5.7 times higher than a patient who at the time of diagnosis was in stage II;
(4) A patient who was overweight at the time of diagnosis has a 20% higher risk of death compared to a normal weight patient.

The assumption inherent in the Cox regression model is that the risk rate or hazard function does not depend on t (time), such that the influence that covariates exert on risk does not change over time. If this were to occur, it would invalidate any analysis and interpretation of the risk inherent in each of the covariates, even if everything else remained constant.

According to the results of the scaled Schoenfeld residual test obtained in Table 2, the p-values are nonsignificant for all covariates, indicating that the residuals are not correlated with time. This confirms that the proportional hazards assumption holds, supporting the conclusion that the risk associated with these variables is time-independent throughout the follow-up period.

It was also verified whether the continuous variable age calculated for the model was on the correct scale, that is, linear, or whether, on the contrary, it needs to be changed. For that purpose, we have analyzed the martingale residuals plot. The results confirmed that the log-hazard relates linearly to age, with no significant evidence of nonlinear patterns, so no changes will be made to the age at diagnosis variable. Therefore, all assumptions were met, that is, linearity and residuals.

Table 2. Results of the scaled Schoenfeld residuals test for the proportional hazards assumption in the Cox model.

Covariate	*Chisq*	*df*	p-value
Age	1.164	1	0.28
StageCAT	2.697	2	0.26
BMICAT	0.798	1	0.37
Global	4.792	4	0.31

Chisq denotes the χ^2 test statistic, with *df* the corresponding degrees of freedom.

3.2. Parametric model

As in Cox's proportional hazards model, the response variable in parametric models is the time from diagnosis to death from breast cancer. The covariates included in the parametric multivariate model are stage, BMI, and age at diagnosis, previously identified as relevant in the Cox model adjustment.

When adjusting Weibull and exponential proportional hazard models, the `survreg` function available in the "`survival`" package in R software is generally used. When the model is written in the form of a proportional hazard model, this function does not directly return the model parameter estimates, so it is necessary to calculate them later.

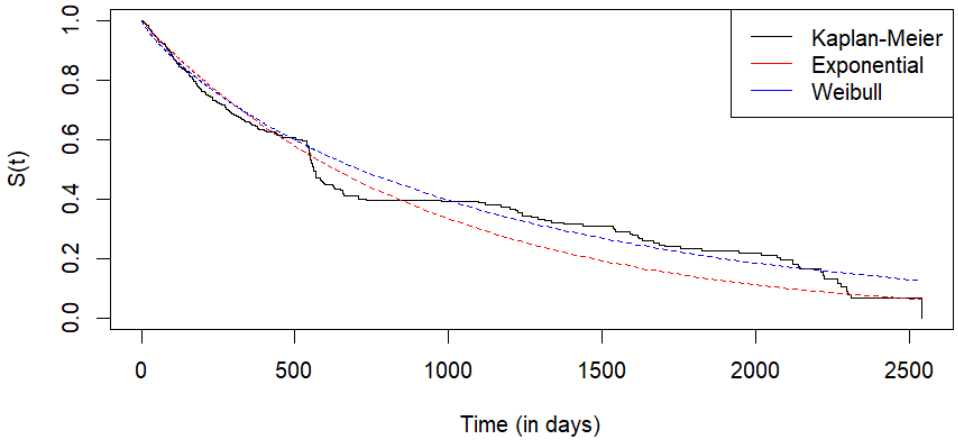

Figure 4. Comparison between the Kaplan–Meier curve and the exponential and Weibull models.

In the analysis of a possible parametric model, the curves presented in Figure 4 were obtained, which show that the distribution that seems to best fit the data, taking into account its distribution around the Kaplan–Meier curve, is the Weibull distribution.

With the aim of comparing parametric models in the form of proportional hazards with semiparametric and flexible approaches, a comparison was made of the parameter estimates of the exponential and Weibull models, as shown in Table 3.

Table 3. Comparison of exponential and Weibull models.

	Exponential (AIC = 6888.24)		Weibull model (AIC = 6862.58)	
	$\widehat{HR}$	p-value	$\widehat{HR}$	p-value
Age	1.03	< 0.001	1.029	< 0.001
Stage (grade III)	2.79	< 0.001	2.814	< 0.001
Stage (grade IV)	5.63	< 0.001	5.623	< 0.001
BMI (overweight)	1.27	0.019	1.238	0.022

In terms of hazard ratio, there are many similarities between the exponential and Weibull models, but Weibull was chosen because it has the lowest Akaike's information criterion (AIC) value. The AIC is a statistical measure that assesses how well the model fits the data, taking into account the model's complexity. It is defined as twice the number of estimated parameters in a model (a measure of complexity) minus twice the maximum value of the model's likelihood function (a measure of goodness of fit). This measure is used for model selection, where a lowest value suggests a more reliable model. Penalizing the number of parameters also ensures a more parsimonious model.

When adjusting the Weibull proportional hazards model with $\alpha = 0.05$, and similarly to the Cox semiparametric model, the only variables that proved significant were age at diagnosis, stage of the disease, and BMI.

Table 4. Comparison of Cox and Weibull models.

	Cox model		Weibull model	
	$\widehat{HR}$	p-value	$\widehat{HR}$	p-value
Age	1.029	< 0.001	1.029	< 0.001
Stage (grade III)	2.871	< 0.001	2.814	< 0.001
Stage (grade IV)	5.684	< 0.001	5.623	< 0.001
BMI (overweight)	1.202	0.049	1.238	0.022

Observing the results in Table 4, it can be seen that the relative risk estimates obtained with the Weibull proportional hazards model are very similar to those obtained with the Cox model.

- The relative risk estimate corresponding to the age translates into an approximately 3% increase in the risk of death for each additional age at the time of diagnosis.

- The relative risk of death for the group of women who were in stage III was estimated to be three times higher compared to a patient who was in stage II or below. For the same covariate, the estimated risk for stage IV indicates that for a patient who was in stage IV at the time of diagnosis, her risk of death increases six times compared to a patient in stage II or below.
- The relative risk for the group of women who were overweight was estimated to be around a 20% higher risk of death compared to a patient who was of normal weight.

3.3. Flexible parametric model

The flexible models proposed by Royston and Parmar can be adjusted in R software using the `flexsurvspline` function from the "`flexsurv`" package or "`Rstpm2`" [15]. The complexity of the model depends on the number of internal knots included in the spline function, defined in the `flexsurvspline` function by the argument k.

The transformation of the survival function in R is done using the scale argument. To obtain the proportional hazards model, the argument scale = "hazard" is used. According to the authors, when no internal knots are included in the proportional hazards model, the simplest model is obtained with only two parameters, which coincides with the Weibull model.

As mentioned earlier, the flexible proportional hazards model is obtained by modelling the logarithm of the cumulative hazard function, $\ln(H(t|\mathbf{x}))$, which can be obtained through the corresponding transformation of the survival function (2.20).

Table 5. Results of the flexible models with up to three internal knots: parameter estimation and AIC value.

	Estimate	Knots			
		0	1	2	3
Age	$\widehat{\beta}_1$	0.028	0.028	0.028	0.028
Stage (grade III)	$\widehat{\beta}_2$	1.035	1.051	1.051	1.051
Stage (grade IV)	$\widehat{\beta}_3$	1.727	1.731	1.731	1.731
BMI (overweight)	$\widehat{\beta}_4$	0.214	0.189	0.189	0.188
AIC	$---$	6862.59	6824.89	6826.74	6828.72

Looking at the values of Table 5, in all the flexible models analyzed, the inclusion of one internal knot in the model decreases the AIC value, thus showing that the inclusion of the knot improves the model's fit.

This fact can be verified by comparing the AIC values for each model, as shown in Table 6. The AIC value for the flexible proportional hazards model with an internal knot is lower than that for the Weibull model, suggesting that it is the better model of the two considered.

Having said this, and in order to analyze a more parsimonious model, a flexible PH model with one internal knot was adjusted, whose estimate of the coefficients, hazard ratio, and respective confidence interval at 95% can be found in Table 7.

Table 6. Comparison of flexible PH models and Weibull PH models: AIC and likelihood ratio test (LRT).

Models	*LRT*	p-value	*AIC*
Flexible PH	6979.8	< 0.001	6824.89
Weibull PH	7019.0	< 0.001	6862.59

Table 7. Flexible proportional hazards model with one internal knot: estimation of coefficients, parameters, hazard ratio, and respective confidence interval at 95% (CI95%).

Covariate	$\widehat{\beta}$	$\widehat{HR}$	*CI*95%
Age	0.028	1.028	$]1.021; 1.036[$
Stage (grade III)	1.051	2.861	$]2.232; 3.668[$
Stage (grade IV)	1.731	5.644	$]4.268; 7.467[$
BMI (overweight)	0.189	1.194	$]1.006; 1.451[$

The interpretation of the estimated relative risks is similar to the one obtained by the Cox model, and the Weibull model, but we can highlight a slight increase in relative risks for older women in more advanced stages of the disease and a very slight decrease in the estimated risk for overweight women when compared with normal weight women.

4. Conclusions

In this research, data was collected at the national level from women diagnosed and treated for breast cancer and followed up, either as inpatients or outpatients, at the Cancer Control Institute of Angola (IACC), to study, analyze, and characterize women fighting breast cancer in the country, not only in descriptive terms through an initial and exploratory analysis of the data but also to infer about this population and detect possible risk and/or protective factors, verifying which aspects affect their survival time.

Several models were applied to describe the survival time of the patients under study from the moment they were diagnosed with the disease until the "death" event occurred. We started by applying the Cox semiparametric model. After validating the assumptions inherent in the model considered and making the necessary adjustments when analyzing the residuals, a model was obtained with the variables that statistically significantly determine the survival time of the patients. These are the stage of the disease at the time of diagnosis, the age at which the disease was diagnosed, and the body mass index at the time of diagnosis.

Subsequently, survival parametric models were fitted to the data. The analysis was carried out by fitting the exponential and Weibull parametric models. Between these two models, the one that best fitted the data was the Weibull parametric model.

When comparing the Cox's semiparametric model with Weibull's model, it can be seen that the relative risk estimates obtained with the two models are quite similar.

Royston and Parmar developed a new family of parametric models with the aim of increasing flexibility compared to existing models, while maintaining the advantages of the Cox model. This model revealed a better adjustment to our data than the previous ones. According to this model, as

expected, a patient in stage III or IV at the time of diagnosis has a higher risk of dying from breast cancer than a patient in stage II, being almost three times higher risk for patients in stage III and almost six times increased risk for those patients in stage IV. For each additional year of age, the risk increases by approximately 3%. Furthermore, a patient who is overweight at the time of diagnosis has 20% higher risk of death than a patient of normal weight.

When there is evidence that a given lifetime distribution is appropriate, it is preferable to use parametric models, as they are more efficient. However, Royston and Parmar models, besides being parametric models, have greater flexibility and are therefore expected to more reliably represent the risk function associated with each covariate.

The flexible proportional hazards model improved the estimate of the survival function, especially when compared to the Weibull model. The flexible proportional proved to be the most appropriate model. In theory, the use of flexible models to analyze the effect of covariates on the time from the diagnosis of the disease to the occurrence of the event results in more accurate estimates of relative risk than using other models, thus contributing to a better understanding of the phenomenon under study.

For all applied models, age, stage of the disease, and body mass index were identified as significant variables to explain the survival time of the population of Angolan women with breast cancer. These are well-known risk factors, and our results confirm their independent effects within this cohort. The lack of statistical significance in other plausible variables, such as treatment or contraceptives, highlights the primacy of these factors within the distinct socioeconomic, demographic, and social stigma realities of Angola. The direction and magnitude of these estimates therefore provide meaningful insights into risk stratification for a population that remains underrepresented in global clinical literature.

It is expected that the results and conclusions drawn from this study will be of some use to professionals whose challenge it is to approach and monitor patients diagnosed with breast cancer, in order to improve their strategies and therefore the health and quality of life of their patients as well as to those responsible for or managing the institutions likely to receive and treat these women.

It is also expected that for those responsible for the development of campaigns to raise awareness and prevent the disease, the set of characteristics identified in the patients studied will allow them to look more closely at the situation in the community where they intend to intervene. Knowledge of the epidemiology of cancer, its determinants, and its risk factors is fundamental for planning and implementing preventive and control interventions.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to thank the editor and three anonymous reviewers for their careful reading of the manuscript and for their constructive comments and suggestions, which helped to improve the quality of the paper. This research was partially supported by Fundação para a Ciência e a Tecnologia, grants UID/04674/2025 (DOI: https://doi.org/10.54499/UID/04674/2025) and UID/315/2025 (DOI:

https://doi.org/10.54499/UID/00315/2025).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Global cancer observatory: Cancer today, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today.
2. H. Moet, Message from the WHO regional director for Africa, world cancer day, Botswana, 2024. Available from: https://www.afro.who.int/regional-director/speeches-messages/world-cancer-day-2024.
3. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, **71** (2021), 209–249. https://doi.org/10.3322/caac.21660
4. D. R. Cox, Regression models and life-tables, *J. R. Stat. Soc. B (Methodol.)*, **34** (1972), 187–202.
5. P. Royston, M. K. B. Parmar, Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects, *Stat. Med.*, **21** (2002), 2175–2197. https://doi.org/10.1002/sim.1203
6. J. P. Klein, H. C. Houwelingen, J. G. Ibrahim, T. H. Scheike, *Handbook of Survival Analysis*, 1st edition, Chapman and Hall/CRC, 2014.
7. N. Younes, J. Lachin, Link-based models for survival data with interval and continuous time censoring, *Biometrics*, **53** (1997), 1199–1211. https://doi.org/10.2307/2533490
8. F. J. Aranda-Ordaz, On two families of transformations to additivity for binary response data, *Biometrika*, **68** (1981), 357–363. https://doi.org/10.1093/biomet/68.2.357
9. J. O. Fernandes, C. Cardoso-Filho, M. B. Kraft, A. S. Detoni, B. N. Duarte, J. Y. Shinzato, et al., Differences in breast cancer survival and stage by age in off-target screening groups: A population-based retrospective study, *AJOG Global Rep.*, **3** (2023), 100208. https://doi.org/10.1016/j.xagr.2023.100208
10. R Core Team, R: A language and environment for statistical computing, 2023. Available from: https://cran.r-project.org.
11. Posit team, RStudio: Integrated development environment for R, PBC, Boston, MA, 2025. Available from: http://www.posit.co.
12. E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, *J. Am. Stat. Assoc.*, **53** (1958), 457–481. https://doi.org/10.1080/01621459.1958.10501452
13. N. Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemother. Rep.*, **50** (1966), 163–170.
14. D. W. Hosmer, S. Lemeshow, S. May, *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, 2nd edition, John Wiley & Sons, Inc., Hoboken, 2008. https://doi.org/10.1002/9780470258019

15. C. Jackson, Flexsurv: A platform for parametric survival modeling in R, *J. Stat. Software*, **70** (2016), 1–33. https://doi.org/10.18637/jss.v070.i08

AIMS Press