https://doi.org/10.48048/tis.2023.6170

Spline Model in The Case of Cervical Cancer Patient Resilience

Rahmat Hidayat[1,*], Muhammad Ilyas[1] and Yuliani[2]

[1]Master of Mathematics Education Program, Cokroaminoto Palopo University, Sulawesi Selatan 91911, Indonesia
[2]Department of Mathematics, Cokroaminoto Palopo University, Sulawesi Selatan 91911, Indonesia

([*]Corresponding author's e-mail: dayatmath@gmail.com)

Received: 3 October 2022, Revised: 8 January 2023, Accepted: 14 March 2023, Published: 21 March 2023

Abstract

Cervical cancer ranks highest in developing countries, and ranks 10th in developed countries as a cause of death for women. In Indonesia, cervical cancer ranks second out of the 10 most common cancers based on Anatomical Pathology data in 2010 with an incidence rate of 20 %. The number of new women with cervical cancer ranges from 90 - 100 cases per 100,000 population and every year there are 40 thousand cases of cervical cancer. Cervical cancer is a gynecological disease that has a high level of malignancy and is caused by the Human Papilloma Virus (HPV). One method that can be used in the analysis of human survival is survival analysis using the Cox proportional hazard model. This research will use the development of the Cox model using the Spline function as its basis. Based on the results obtained, the developed model can model the data well and give a low MSE. Based on the developed model, it is found that the variables that influence the survival rate of cervical cancer patients are age, stage, type of PRC transfusion treatment and comorbidities.

Keywords: Cancer, Cervical, Cox, Model, Spline, Survival, Women

Introduction

Cervical cancer is the growth of abnormal cells in the cervical tissue (cervix) and is a primary cancer originating from the cervix. The cervix is the part of the front end of the uterus that protrudes into the vagina. Several factors that can increase the risk of cervical cancer according to [1], include sexual intercourse, history of pregnancy, Diethylstilbesterol (DES), infectious agents, and history of smoking.

Cells don't just turn into cancer. Normal cells in the cervix gradually turn from pre-cancerous to cancerous. Cervical cancer begins when normal cells on the surface of the cervix change and grow uncontrollably to form a mass known as a tumor. Tumors can be cancerous or benign. Malignant tumors can spread to other parts of the body, whereas benign tumors cannot. Doctors use several measures to describe precancerous changes, including cervical intraepithelial neoplasia (CIN), squamous intraepithelial lesion (SIL), and dysplasia that can be detected by a Pap smear. Cervical cancer consists of 2 types, squamous cell carcinoma and adenocarcinoma. About 80 - 90 % of cervical cancer is squamous cell carcinoma. This cancer is formed from cells that are in the exocervix. Another type of cancer is adenocarcinoma, which forms from gland cells. Cervical cancer and pre-cervical cancer are caused by the human papilloma virus (HPV). HPV is a collection of more than 150 related viruses. Different types of HPV cause damage (warts) on different parts of the body. HPV types with low risk (cervical cancer) are HPV 6 and HPV 11, because these types are rarely associated with cervical cancer.

One of the models in statistics that can be used to analyze a person's survival is survival analysis. Survival analysis is one of the methods in statistics where the variable you want to look at is time, until the occurrence of an event. In this case the incident in question is a relapse of an illness, disease outbreak, death, healing and other events that can happen to a person. Survival analysis assumes that there is only one event, although in fact there may be more than one event in the same study [2]. In analyzing survival data, ordinary linear regression cannot be used because ordinary linear regression is not able to handle censored observations, namely observations that are not observed because they are missing or incomplete.

Several methods used to analyze the relationship between predictor variables and survival time include parametric, nonparametric and semiparametric methods. The parametric method assumes that the distribution underpinning the survival time follows a certain distribution such as exponential, gamma, Weibull and so on. If the underlying distribution of survival time is unknown, meaning that the data does

not follow a certain pre-existing distribution, then the nonparametric method is used [3]. There are 2 well-known nonparametric methods, namely the Kaplan-Meier and Nelson-Aalen methods. However, Kaplan-Meier is used more often than Nelson-Aalen. Meanwhile, if the survival data to be studied involves many predictor variables, then regression can be used.

The regressions used to analyze the survival data are parametric regression, nonparametric regression, and semiparametric regression. Parametric regression requires the condition that the baseline hazard follows a certain distribution. If these conditions are not met, nonparametric regression can be used. Nonparametric regression is used if the data used does not follow a certain pre-existing distribution. Several nonparametric regression methods are spline, kernel, Fourier series and (Multivariate Adaptive Regression Splines) MARS. Spline was first introduced by Whitaker in 1923 as a data pattern approach. Spline based on an optimization problem was developed by Reinsc in 1967 [4]. The spline approach has a basic function, commonly used are truncated splines and B-splines. The truncated spline is a function where there is a change in the behavior pattern of different curves at different intervals. The advantage of truncated splines is that they can describe changes in the behavior pattern of the function at certain sub-intervals. Meanwhile, [5-7] examined the determination of knot points in spline regression.

[8-10] in their research conclude that the MSE value on the truncated spline curve is smaller than the linear regression on all functions, this means that the truncated spline curve is better than the linear regression. Survival analysis with regression has been developed by [11], regarding modeling relapse time in cases of breast cancer patients where the effect of predictor variables changes at the time of observation. Some of the predictor variables were lymph nodes, tumor size, age, time of menopause, and weight of breast cancer patients. On the interaction between lymph nodes and tumor size using splines in the first group having 1 - 2 positive lymph nodes and the second group having 15 significant positive lymph nodes where patients who have many positive nodes tend to have metastatic disease at the start of the study, regardless of tumor size. primary and tends to have a rapid recurrence in the presence of metastases. Meanwhile, in patients with few lymph nodes, the spread of metastases tends to be less and the size of the primary tumor is more influential on the time of recurrence.

In conducting survival analysis, the basic function used is a linear regression model. This model assumes that risk factors are linear. In actual fact, sometimes obtained data with risk factors that are not in accordance with the linear function. So that in this study, the existing survival model was developed by changing its basic function to a non-linear or non-parametric function. This research will conduct a case study on cervical cancer patients at the Regional General Hospital Dr. Soetomo Surabaya, Indonesia. Cervical cancer ranks highest in developing countries, and ranks 10th in developed countries. In Indonesia, cervical cancer ranks second out of the 10 most cancers based on data from Anatomical Pathology in 2010 with an incidence of 20 %. According to current estimates from the Ministry of Health of the Republic of Indonesia (RI), the number of new women with cervical cancer ranges from 90 - 100 cases per 100,000 population and every year there are 40 thousand cases of cervical cancer. This study aims to use the spline regression model for survival analysis with residual cox PH in cervical cancer patients.

Materials and methods

Survival Function and Hazard function

In the survival analysis, there are 2 main functions, namely the survival function and the hazard function. Survival function serves to determine the probability of the patient's survival time from the start point to time *t*. *T* is the notation of survival time and is a random variable that has a probability distribution function *f(t)*, then the probability density function can be expressed as follows [12,13].

$$f(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t+\Delta t)}{\Delta t} \quad (1)$$

The cumulative distribution function can be expressed as follows.

$$F(t) = P(T \leq t) = \int_0^t f(t)dt \quad (2)$$

The survival function *S(t)*, defined as the probability of an object surviving after the *t*-th time, is expressed by the following equation.

$$S(t) = P(T > t) = 1 - F(t) = 1 - P(T \leq t) \quad (3)$$

The Hazard function is a function that states the instantaneous failure rate when experiencing an event at the *t*-th time or it can be said that the individual chance of experiencing an event in the *t*-th time, the equation of the Hazard function can be stated as follows.

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{p(t \le T < t + \Delta t \mid T > t}{\Delta t} \right\} \quad (4)$$

Hazard function, which states the rate of failure of an individual to experience an event in the time interval from *t* to $t + \Delta t$ provided that an individual has survived until time t. For example, the probability that the random variable *T* is greater than or equal to *t* is between *t* and $t + \Delta t$, provided that *t* and *T* are greater than or equal to *t*. Based on the equation, the relationship between the Survival function and the Hazard function can be obtained using the conditional probability theory as follows. P(A∩B) is a probability of a joint event between A and B. The conditional probability theory can be formulated as follows [14].

$$P(\mathrm{A} \mid \mathrm{B}) = \frac{P(A \cap B)}{P(B)} \quad (5)$$

Suppose *f(t)* is the probability density function at time t, then from the above equation it is obtained as follows [15].

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \left\{ \frac{p(t \le T < t + \Delta t | T \ge t)}{\Delta t} \right\} \\ &= \lim_{\Delta t \to 0} \left\{ \frac{P\left(t \le T < (t + \Delta t) \cap (T \ge t)\right)}{\Delta t \times P(T \ge t)} \right\} \\ &= \lim_{\Delta t \to 0} \left\{ \frac{P(t \le T < (t + \Delta t))}{\Delta t \times S(t)} \right\} \\ &= \frac{1}{S(t)} \times \lim_{\Delta t \to 0} \left\{ \frac{P(t \le T < (t + \Delta t))}{\Delta t} \right\} \end{aligned}$$

Thus, the relationship between the Survival function and the Hazard function can be stated as follows [16].

$$h(t) = \frac{f(t)}{S(t)} \quad (6)$$

Based on Eqs. (6) - (8) can be obtained as follows.

$$f(t) = \frac{d\left(F(t)\right)}{dt} = \frac{d(1 - S(t))}{dt} = \frac{d(S(t))}{dt} \quad (7)$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{dS(t)}{dt}, \frac{d\ln S(t)}{d.S(t)} = \frac{d\ln S(t)}{dt} \quad (8)$$

Based on Eq. (8), the following equation can be obtained.

$$\int_0^t h(t)dt = -\int_0^t \frac{d\ln S(t)}{dt} dt = -\int_0^t \frac{d}{dt} \ln Sit) dt$$

so,

$$-\int_0^t h(t)dt = \ln S(t) \mid_0^t = \ln S(t) - \ln S(0)$$

$$-H(t) = \ln S(t)$$

so, the survival function can be formulated as follows;

$$\mathrm{S(t)} = \exp\left(-H(t)\right) \quad (9)$$

where the cumulative hazard function is as follows.

$$H(t) = \int_0^t h(t)dt \tag{10}$$

The function *H(t)* is the cumulative Hazard function obtained from the Survival function. Based on Eqs. (9) - (10), it can be obtained the relationship between the Survival function and the cumulative Hazard function as follows [17].

$$H(t) = -\ln S(t) \tag{11}$$

Spline function

Polynomial slices play an important role in approximation theory and statistics. Polynomial slices are flexible and effective in dealing with local properties of a function or data [18-20]. One of the most important types of polynomial slices is the spline polynomial. The spline as a data pattern approach was adopted by Whittaker in 1923, while the spline based on an optimization problem was developed by Reinsch in 1967 [21-24].

Spline polynomial of order *m* with knot points $\lambda_1, \lambda_2, \cdots, \lambda_r \ (a < \lambda_1 < \ldots < \lambda_r < b)$ is a function *f* which is expressed in the form;

$$f(x) = \sum_{i=0}^{m} \theta_i x^i + \sum_{j=1}^{r} \phi_j \left(x - \rho_j\right)_+^m, \tag{12}$$

with $\theta_i, i = 0,1,2, \ldots, m$ and $\emptyset_j, j = 1,2, \ldots, r$ are real-valued constants, and;

$$\left(x - \lambda_j\right)_+^m = \begin{cases} \left(x - \lambda_j\right)^m, \text{if } x \geq \lambda_j \\ 0 \qquad\quad , \text{if } x < \lambda_j \end{cases}$$

Spline nonparametric regression

Many studies of spline estimators on nonparametric regression models have been carried out such as [25-29]. The nonparametric regression model which states the relationship between the predictor variable and the response variable can be written as follows;

$$y_i = f(x_i) + \varepsilon_i, \ i = 1,2,\cdots,n, \tag{13}$$

Eq. (12) can be changed in the form of matrix multiplication;

$$\mathbf{f}(x) = \mathbf{X}(\boldsymbol{\lambda})\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{14}$$

with $\mathbf{X}(\boldsymbol{\lambda})$ is a matrix that depends on $\boldsymbol{\lambda}$, whereas $\boldsymbol{\beta}$ is a vector that depends on $\boldsymbol{\theta}$, and ϕ. To get an estimator of *f*, the Maximum Likelihood Estimation (MLE) method is also used. The first estimate is to estimate $\boldsymbol{\beta}$, with likelihood function;

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\lambda}) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} Exp\left(-\frac{\varepsilon_i^2}{2\sigma^2} \right) \right) \\ &= (2\pi\sigma^2)^{-n/2} Exp\left(-\frac{1}{2\sigma^2} \| \boldsymbol{\varepsilon} \|^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} Exp\left(-\frac{1}{2\sigma^2} \| \mathbf{y} - \mathbf{X}(\boldsymbol{\lambda})\boldsymbol{\beta} \|^2 \right) \end{aligned}$$

so that obtained;

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \left[\left(\mathbf{X}(\boldsymbol{\lambda}) \right)' \mathbf{X}(\boldsymbol{\lambda}) \right]^{-1} \left(\mathbf{X}(\boldsymbol{\lambda}) \right)' \mathbf{y}$$

causes an estimate for $\mathbf{f}(x)$ is;

$$\hat{\mathbf{f}}(x) = \mathbf{X}(\boldsymbol{\lambda})\hat{\boldsymbol{\beta}} = \mathbf{X}(\boldsymbol{\lambda}) \left[\left(\mathbf{X}(\boldsymbol{\lambda}) \right)' \mathbf{X}(\boldsymbol{\lambda}) \right]^{-1} \left(\mathbf{X}(\boldsymbol{\lambda}) \right)' \mathbf{y}$$

The knot points are the controller of the balance between the smoothness of the curve and the suitability of the curve to the data. The selection of knot points can be done using the Generalized Cross Validation or GCV method.

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{\mathrm{MSE}(\boldsymbol{\lambda})}{\left(n^{-1} tr(\mathrm{I} - \mathrm{S}(\boldsymbol{\lambda})) \right)^2}, \tag{15}$$

where $\mathrm{MSE}(\boldsymbol{\lambda})$ on the Eq. (15) is;

$$\mathrm{MSE}(\boldsymbol{\lambda}) = n^{-1} \sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2$$

The selection of optimal knot points is done by looking at the smallest GCV value.

Results and discussion

Characteristics of cervical cancer patients

This section discusses the characteristics of cervical cancer patients based on survival time and the factors that are thought to affect the survival of cervical cancer patients who are hospitalized at RSUD Dr. Soetomo Surabaya Indonesia. Descriptive statistical analysis aims to determine the characteristics of the patient. Characteristics of cervical cancer patients based on survival time and factors that are thought to affect the survival of cervical cancer patients who are hospitalized at RSUD Dr. Soetomo Surabaya.

Table 1 Descriptive statistics by stadium.

Variable	D	Stadium	N	Mean	Min	Max
Survival Time	Censor	0	15	24.51	2	226
		1	27	14.68	2	58
		2	225	47.79	1	331
		3	501	48.03	1	329
		4	9	71.33	2	243
	Dead	0	1	5	5	5
		1	0	*	*	*
		2	2	18	18	18
		3	21	38.71	1	151
		4	16	8.32	1	43

Table 1 shows the descriptive statistics of patient survival time by stage level. At the stage 0 level, there were 15 patients who remained in treatment for 24 days. At the stage 1 level there were 27 patients who remained in treatment for 14 days. At the stage 2 level, there were 225 patients who could survive 48 days of treatment. Stage 3 is the largest number of patients, namely 501 patients who survived in treatment with an average survival time of 48 days. Meanwhile, stage 4 was the smallest number, namely 9 patients who were able to survive in treatment with an average survival time of 71 days. At the end of the study, there was 1 patient who died at stage 0 with a survival time of 5 days. Stage 1 no patient died. In Stage 2, 2 patients died with a median survival time of 18 days. In Stage 3, 21 patients died with an average survival time of 38 days. In Stage 4, 16 patients died with an average survival time of 8 days.

Table 2 Descriptive statistics by type of treatment.

Variable	D	Stadium	N	Mean	Min	Max
Survival Time	Censor	Chemotherapy	411	46.53	1	332
		PRC Transfusion	244	46.87	2	278
		PRC Chemotherapy and Transfusion	60	55.4	3	338
		Operation	72	45.11	1	295
	Dead	Chemotherapy	8	22	1	94
		PRC Transfusion	19	20.13	1	145
		PRC Chemotherapy and Transfusion	7	49.77	3	154
		Operation	5	10.8	4	17

Table 2 is a descriptive statistical table of patient survival time by type of treatment. In the type of PRC transfusion treatment, there were 244 patients who survived during treatment with an average survival time of 47 days. In the type of chemotherapy treatment and PRC transfusion, there were 60 patients who survived during treatment with an average survival time of 55 days. Meanwhile, the smallest number of patients was in the type of surgical treatment as many as 72 patients who survived during treatment with an average survival time of 45 days. The type of chemotherapy treatment was the largest number of patients, namely 411 patients who survived during treatment with an average survival time of 46 days.

At the end of the study, there were 9 patients who died with this type of chemotherapy treatment with an average survival time of 22 days. In the type of PRC transfusion treatment, there were 19 patients who died with an average survival time of 20 days. Patients who underwent chemotherapy treatment and PRC transfusion died as many as 7 patients died with an average survival time of 50 days. Patients who underwent surgical treatment died as many as 5 patients with an average survival time of 11 days.

Based on the type of treatment, there are 3 types of treatment undertaken by cervical cancer patients based on the doctor's decision or medical personnel seen from the patient's condition before undergoing treatment. Patients who survived by doing this type of chemotherapy treatment as many as 472 people. Patients who survived by doing this type of PRC transfusion treatment were 296 people. Patients who survive by doing this type of surgical treatment as many as 70 people. In addition, it is also known that 16 patients died by undergoing chemotherapy treatment. There were 26 patients who died using PRC transfusion treatment. There were 5 patients who died by undergoing surgical treatment. Of the 3 types of treatment, it is known that chemotherapy is the best type of treatment because it has the greatest probability of survival for cervical cancer patients. Meanwhile, the type of PRC transfusion treatment has the lowest probability of survival compared to other types of treatment.

Based on the data obtained, it is also known that the patient not only has the main disease, namely cervical cancer, but the patient also has other comorbidities such as pulmonary edema, anuria, obesity, and hyperglycemia. Patients who survived and had comorbidities were 41 patients. Patients who survived and did not have comorbidities were 670 patients. There were 6 patients who died and had co-morbidities. Patients who died and did not have comorbidities were 30 patients. Based on the description above, it is known that patients who have comorbidities have a lower probability of survival than patients who do not have comorbidities. As for the complication variable, it can be seen that patients who experience complications have a lower probability of survival compared to patients who do not experience complications.

Based on the patient's stage, it was known that in stage 0 there were 15 patients who survived and 1 patient died, stage 1 there were 27 patients who survived and no patient died. In Stage 2, 225 patients survived and 2 patients died. Stage 3 there were 501 patients who survived and 21 patients who died. Stage 4 there are 9 patients who survive and 16 patients who died. Based on the description above, it can be concluded that patients who are at stage 4 have the lowest probability of survival compared to patients who are at other stages. Meanwhile, patients in stage 1 have the lowest probability of survival disease compared to patients at other stages.

The last variable is anemia. Patients who survived did not experience anemia as many as 286 patients and patients who survived experienced anemia as many as 455 patients. Meanwhile, in contrast to the condition of patients who died, 29 patients had anemia and 11 patients died without anemia. Cervical cancer patients who are anemic will experience acute or chronic bleeding. Cervical cancer patients who experience anemia have a lower probability of survival than cervical cancer patients who do not experience anemia.

Cox proportional hazard model for cervical cancer patients

In this section, we will discuss the cox proportional hazard model of cervical cancer patients at Dr. RSUD. Soetomo Surabaya based on the factors that are thought to affect the survival of cervical cancer patients who are hospitalized at RSUD Dr. Soetomo Surabaya. Kaplan Meier Survival Curve Analysis and Log Rank Test Characteristics of the survival time of cervical cancer patients can be shown using the Kaplan Meier survival curve. Meanwhile, to find out whether there is a difference between the survival curves of different factor groups, the Log Rank test can be used. The following is Kaplan Meier's overall survival curve to find out the general characteristic description given in **Figure 1** below.

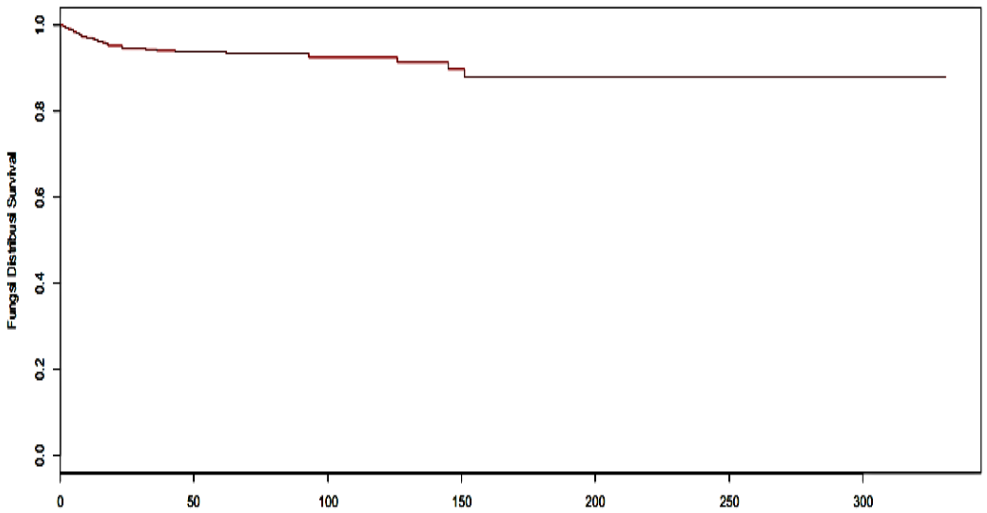

Figure 1 Kaplan Meier survival curve for cervical cancer patients.

Figure 1 shows that on day 0 to day 331 the survival curve decreases slowly. The probability of survival of cervical cancer patients in the time range from day 0 to day 150 is 0.88. This value shows the probability of survival of cervical cancer patients is still high. Meanwhile, the curve on the 150th day of survival probability of cervical cancer patients tends to be constant. **Figure 4** is a general description of the characteristics of the survival curve. The following will describe the characteristics of the survival curve based on the factors that are thought to affect the survival of cervical cancer patients.

Next will be seen the characteristics of each variable used. To see the difference in the survival curve, statistical testing can be done by using the Log Rank test. To test the hypothesis whether there is a difference between the survival curves of cervical cancer patients based on the factors thought to affect the survival of cervical cancer patients, it is necessary to perform a log rank test on each of the variables presented in **Table 3**.

Table 3 Log rank test results for each variable.

Variable	Log rank	*p*-value
Age	77.6	0.0074
Stadium	272	0
Chemotherapy	8.5	0.0036
PRC Transfusion	12.3	0.0005
Operation	0.8	0.378
co-morbidities	4.6	0.00327
Complications	77.5	0
Anemia status	9.7	0.0018

Based on **Table 3**, it can be concluded that the significance level $\alpha = 0.05$, the variables of age, stage, type of chemotherapy treatment, type of PRC transfusion treatment, comorbidities, complications, and anemia status have different Kaplan Meier survival curves.

Testing the proportional hazard assumption

The proportional hazard assumption test is carried out before constructing the model. It aims to determine the factors that meet the proportional hazard assumption. Testing the proportional hazard assumption using the goodness of fit method.

Table 4 Goodness of fit test results.

Variable	Correlation	*p*-value	Decision
Age	0.0086	0.9393	Failed to reject H_0
Stadium	−0.1289	0.0878	Failed to reject H_0
Chemotherapy	0.0931	0.5021	Failed to reject H_0
PRC Transfusion	0.0073	0.921	Failed to reject H_0
Operation	0.1115	0.4372	Failed to reject H_0
co-morbidities	0.1197	0.231	Failed to reject H_0
Complications	−0.0074	0.3821	Failed to reject H_0
Anemia status		0.9214	Failed to reject H_0

Table 4 is the result of goodness of fit analysis on the independent variables using $\alpha = 0.05$, so it can be seen that all the factors that are thought to affect the survival of cervical cancer patients have a *p*-value greater than the value of so it can be concluded that all the factors suspected affecting the survival of cervical cancer patients, namely age, type of chemotherapy treatment, PRC transfusion, surgery, comorbidities, complications and anemia status meet the proportional hazard assumption.

Survival function and cumulative hazard function

The following will describe the curves of the survival function and the hazard function. The survival function is used to determine the probability of survival of cervical cancer patients.

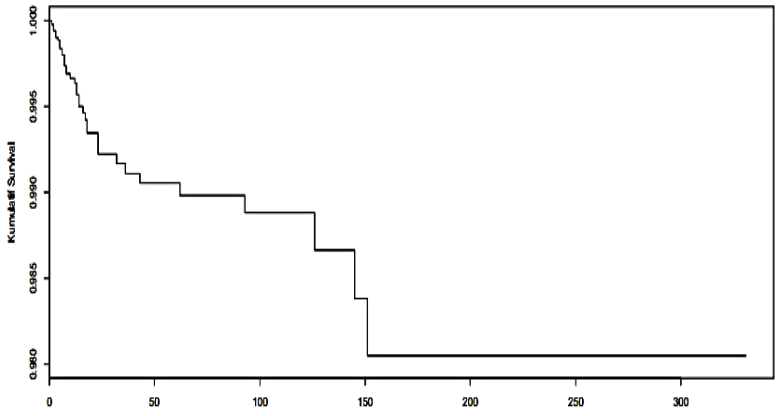

Figure 2 Survival function curve.

The survival function curve expresses the probability of cervical cancer patients surviving at a certain time. The survival function curve from day 0 to day 150 decreased, which means that the probability of cervical cancer was getting smaller on the 0 to 150 day survival time. The 150th day survival function curve of the probability of cervical cancer patients is constant.

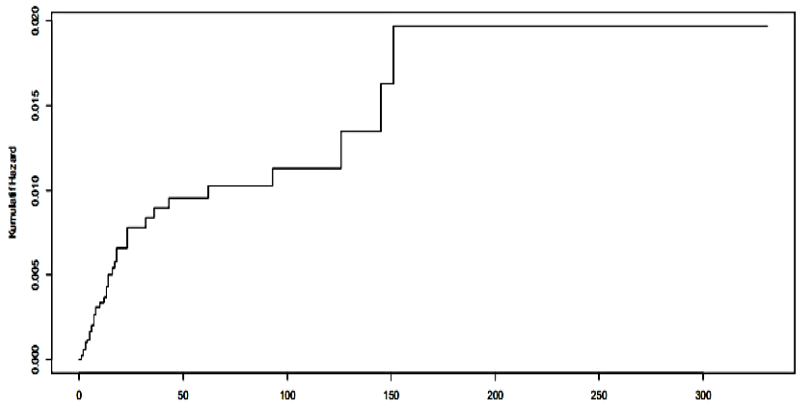

Figure 3 Hazard function curve.

The hazard function curve represents the rate of cervical cancer patients dying. The rate of cervical cancer patients who died on day 0 to day 150 has increased. The rate of cervical cancer patients on the 150th day tends to be constant, meaning that the rate of cervical cancer patients dying is getting higher, and will increase on the 300th day from before.

Cox proportional hazard model

The following is an estimate of the obtained Cox proportional hazard regression, which is presented in **Table 5**.

Table 5 Cox proportional hazard model test results.

Parameter	Parameter Estimation	Z	*p*-value	Decision
Age	9.57×10^{-3}	0.43	0.6703	Failed to reject H_0
Stage 1	−16.5	0	0.9965	Failed to reject H_0
Stage 2	−2.28	−1.84	0.0665	Failed to reject H_0
Stage 3	−1.18	−1.14	0.2556	Failed to reject H_0
Stage 4	1.69	1.60	0.1095	Failed to reject H_0
Chemotherapy 1	0.779	1.62	0.1051	Failed to reject H_0
PRC Transfusion 1	1.94	2.63	0.0086	Reject H_0
Operation 1	2.17	2.75	0.0060	Reject H_0
Accompanying poet 1	0.44	1.08	0.2796	Failed to reject H_0
Complication 1	2.46	4.36	1.3×10^{-5}	Reject H_0
Anemia Status 1	−0.878	−1.60	0.1088	Failed to reject H_0
Likelihood ratio test		130	0	Reject H_0

Based on the parameter results, the Cox proportional hazard regression model is obtained as follows;

$$\hat{h}(t) = h_0(t)exp\Big(9{,}57\times10^{-3}(Age) - 16{,}5\big(Stadium(1)\big) - 2{,}28\big(Stadium(2)\big) - 1{,}18\big(Stadium(3)\big) + 1{,}69\big(Stadium(4)\big) + 0{,}779\big(Chemotherapy(1)\big) + 1{,}94\big(\text{PRC Transfusion }(1)\big) + 2{,}17\big(Operation(1)\big) + 0{,}44\big(Accompanying\ poet(1)\big) + 2{,}46\big(Complication(1)\big) - 0{,}878\big(Anemia\ Status\ (1)\big)\Big)$$

Hazard ratio

The cox proportional hazard model that is formed will interpret the hazard ratio value of each variable used in the modeling which is presented in **Table 6**.

Table 6 *Hazard ratio* model *Cox proportional hazard.*

Parameter	Hazard ratio
Age	1.21
Stage 1	5.33×10^{-7}
Stage 2	0.202
Stage 3	0.314
Stage 4	5.77
Chemotherapy	3.19
PRC Transfusion	5.65
Operation	7.89
Co-morbidities	1.55
Complications	12.51
Anemia Status	0.471

Based on **Table 6**, it can be seen that the hazard ratio value for age is 1.21, meaning that cervical cancer patients who are more than 45 years old have a 1.21 times greater risk of death than patients who are less than 45 years old. The hazard ratio value for patients at stage (1); is 5.33×10^{-7}, which means that when compared to patients at stage 0, patients at stage (1) have a risk of dying of 5.33×10^{-7} times. The hazard ratio value for patients at stage (2); is 0.202, which means that when compared to patients at stage 0, patients in stage (2); have a 0.202 times risk of dying. The hazard ratio value for patients at stage (3); is 0.314, which means that when compared to patients at stage 0, patients at stage (3); have a 0.314 times risk of dying. The hazard ratio value for patients at stage (4); is 5.77, which means that when compared to patients at stage 0, patients at stage (4); have a 5.77 times risk of dying.

The hazard ratio value for patients undergoing chemotherapy treatment (1); is 3.19, which means that when compared with patients not undergoing chemotherapy treatment, patients undergoing chemotherapy (1); have a 3.19 times risk of dying. The hazard ratio value for patients who underwent PRC transfusion (1); was 5.65, which means that when compared with patients who did not undergo any type of PRC transfusion treatment, patients who underwent PRC transfusion had a 5.65 times risk of dying. The hazard ratio value for patients undergoing surgical treatment is 7.89, which means that when compared to patients not undergoing surgical treatment, patients undergoing surgical treatment have a risk of dying of 7.89 times.

The hazard ratio value for patients who have comorbidities is 1.55 which means that when compared to patients who do not have comorbidities, patients who have comorbidities have a risk of dying of 1.55 times. The hazard ratio value for patients experiencing complications is 12.51, which means that when compared with patients who do not experience complications, patients who experience complications have a risk of dying of 12.51 times. The hazard ratio value for patients who have anemia is 0.471, which means that when compared to patients without anemia, patients with anemia have a risk of dying of 0.471 times.

Cox proportional hazard modeling in spline regression

The spline regression model was applied in this study, to model the factors thought to affect the survival of cervical cancer patients. The method used to model this research is spline truncated regression. In this study, the response variables used were residuals from cox PH (martingale residual), while the predictor variables were the variables used in the medical record data of cervical cancer patients. The formation of the spline regression model is by selecting the optimal knot point. The knot point is the point where the pattern of the data changes. To get the optimal knot point, the GCV method is used. To choose the optimal knot value, the minimum GCV value is used. The knot point used in this study is a combination of knots. Selection of optimal knot points with a combination of knots. Knot combination is a combination of 1 knot point, 2 knot point, and 3 knot point. This combination is used to select the optimal knot point. The choice of this knot combination is used because the predictor variable contains continuous and categorical data, where the determination of the knot point will vary. By using a combination of knots on the variables that affect the survival rate of cervical cancer patients will get the minimum GCV value that produces the best spline model. The best truncated spline regression model is obtained from the optimal knot points. To get the optimal knot point, the minimum GCV is used. The following are the results of the GCV calculation for the spline regression of the knot combination.

Table 7 Optimal knot point selection with knot combination.

Variable	Knot point variation	Knot point	GCV
x_1	3	$K_{11} = 41.86;\ K_{12} = 41.86;\ K_{13} = 41.86$	1.356
x_2	3	$K_{21} = 1;\ K_{22} = 2;\ K_{23} = 3$	
x_3	1	$K_{31} = 1$	
x_4	1	$K_{41} = 0$	
x_5	1	$K_{51} = 1$	
x_6	1	$K_{61} = 1$	
x_7	1	$K_{71} = 0$	
x_8	1	$K_{81} = 1$	

Cox proportional hazard model with spline regression

After obtaining the best spline regression model, the model will be substituted in the general cox PH model. The cox PH model assumes $h_0(t)$ is unknown and is not affected by time so that the value is constant.

Cox PH modeling in the spline regression model is as follows;

$$h(t) = h_0(t)exp(0.21606 - 0.00514x_1 - 0.00201(x_1 - 41.86)_+ + 0.016134(x_1 - 49.29)_+ - 0.00823(x_1 - 56.71)_+ - 0.006315x_2 - 0.00321(x_2 - 1)_+ - 0.00153(x_2 - 2)_+ - 0.00779(x_2 - 3)_+ - 0.0032x_3 - 1.28 \times 10^{-17}(x_3 - 1)_+ - 0.00065x_4 - 0.00065(x_4)_+ + 0.000702x_5 - 0.0004x_7 - 0.0004(x_4)_+ + 0.000575x_8)$$

This truncated spline model with a combination of knots has an MSE of 0.05.

Discussion

Based on the results of the analysis of the Cox PH model in the best spline regression with the combination of knots obtained, several interpretations can be explained, namely in patients aged less than 41.86 years, if in this group, the patient's age increases by 1 year, the risk of death will decrease to 0.00514. In patients aged between 41.86 years and 49.29 years, if in this group, the patient's age increased by 1 year, the risk of death would decrease to 0.00715. In patients aged between 49.29 years and 56.71 years, if this group, the patient's age increases by 1 year, the risk of death will increase by 0.008984. In patients aged more than 56.71 years, if this group, the patient's age increases by 1 year, the risk of death will increase by 0.000754.

From this model, it can be interpreted that if the patient is at stage 0 it has the equation $\hat{y} = 0.21606 - 0.006315x_2$. If the patient is in stages between 1 and 2 have the same $\hat{y} = 0.219274 + 0{,}003105x_2$. Furthermore, if the patient is at a stage between 2 and 3 has the same $\hat{y} = 0.222334 + 0.001575x_2$. If the patient is at stage 4 has the same $\hat{y} = 0.245704 - 0.006215x_2$. For the age variable, the model can be interpreted, namely if the patient does not undergo chemotherapy and the patient undergoes the type of chemotherapy treatment has the same $\hat{y} = 0.216064 + 0.0032x_3$.

If the variable $x_2, x_3, x_4, x_5, x_6, x_7, x_8$ considered constant, then the influence of the age variable $x_1$is large on the residual martingale survival rate of cervical cancer patients y is;

$$\hat{y} = 0.21606 - 0.00514x_1 - 0.00201(x_1 - 41.86)_+ + 0.01613(x_1 - 49.29)_+ + \ -0.00823(x_1 - 56.71)_+$$

$$= \begin{cases} 0{,}21606 - 0{,}00514x_1 & x_1 < 41.86 \\ 0{,}300203 - 0.00715x_1 & 41.86 \le x_1 < 49.29 \\ 0.008984x_1 - 0.49504 & 49.29 \le x_1 < 56.71 \\ 0.000754x_1 - 0.02832 & x_1 \ge 56.71 \end{cases}$$

From this model can be interpreted as follows;

1) In patients aged less than 41.86 years, if in this group, the patient's age increases by 1 year, the risk of death will decrease to 0.00514.

2) In patients aged between 41.86 years and 49.29 years if in this group, the patient's age increases by 1 year, the risk of death will decrease to 0.00715.

3) In patients aged between 49.29 years and 56.71 years if in this group, the patient's age increases by 1 year, the risk of death will increase by 0.008984.

4) In patients aged more than 56.71 years, if in this group, the patient's age increases by 1 year, the risk of death will increase by 0.000754.

If the variables $x_1, x_3, x_4, x_5, x_6, x_7, x_8$ considered constant, the influence of the age variable x_2 is large on the residual martingale survival rate of cervical cancer patients y is;

$$\hat{y} = 0.21606 - 0.006315x_1 - 0.00321(x_2 - 1)_+ - 0.00153(x_2 - 2)_+ + -0.00779(x_2 - 3)_+$$

$$= \begin{cases} 0{,}21606 - 0{,}006315x_2 & x_2 < 1 \\ 0{,}300203 - 0.00715x_1 & 1 \le x_2 < 2 \\ 0.008984x_1 - 0.49504 & 2 \le x_2 < 3 \\ 0.000754x_1 - 0.02832 & x_2 \ge 3 \end{cases}$$

From this model it can be interpreted that if the patient is in stage 0, then it has equation $\hat{y} = 0.21606 - 0.00631x_2$. If the patient is in stages between 1 and 2 have equation $\hat{y} = 0.003105x_2 + 0.2192$. Furthermore, if the patient is in stages between 2 and 3 have equation $\hat{y} = 0.22234 + 0.001575x_2$. If the patient is in stage 4, then the equation is $\hat{y} = -0.006215x_2 + 0.245701$.

If the variable $x_1, x_2, x_4, x_5, x_6, x_7, x_8$ considered constant, the influence of the variable x_3 is large on the residual martingale survival rate of cervical cancer patients y is;

$$\hat{y} = 0.21606 - 0.0032x_3 - 1.28 \times 10^{-17}(x_3 - 1)_+$$

$$= \begin{cases} 0.21606 - 0.0032x_3; & x_3 < 1 \\ 0.21606 - 0.0032x_3; & x_3 \ge 1 \end{cases}$$

From this model it can be interpreted that if the patient does not undergo chemotherapy and the patient undergoing this type of chemotherapy treatment has equation $\hat{y} = 0.21606 - 0.0032x_3$

If the variable $x_1, x_2, x_3, x_5, x_6, x_7, x_8$ considered constant, the influence of the variable x_4 is large on the residual martingale survival rate of cervical cancer patients y is;

$$\hat{y} = 0.21606 - 0.00065x_4 - 0.00065(x_4)_+$$

$$= \begin{cases} 0.21606 - 0.00065x_4; & x_4 < 1 \\ 0.21606 & x_4 \ge 1 \end{cases}$$

From this model it can be interpreted that if the patient does not undergo this type of treatment, the PRC transfusion has equation $\hat{y} = 0.21606 - 0.00065x_4$. Furthermore, if the patient undergoes this type of treatment, the PRC transfusion has equation $\hat{y} = 0.21606$.

If the variable $x_1, x_2, x_3, x_4, x_6, x_7, x_8$ considered constant, the influence of the variable x_5 is on the residual martingale survival rate of cervical cancer patients y is;

$$\hat{y} = 0.21606 - 0.000702x_5$$

$$= \begin{cases} 0.21606 - 0.00072x_5; & x_5 < 1 \\ 0.21606 - 0.00072x_5 & x_5 \ge 1 \end{cases}$$

From this model it can be interpreted that if patients who do not undergo surgery and patients undergoing surgery have the same type of treatment $\hat{y} = 0.21606 + 0.00072x_5$

If the variable $x_1, x_2, x_3, x_4, x_5, x_7, x_8$ considered constant, the influence of the variable x_6 is on the residual martingale survival rate of cervical cancer patients y is;

$$\hat{y} = 0.21606 - 0.00038x_6$$

From this model it can be interpreted that if the patient does not have co-morbidities and has co-morbidities then the equation $\hat{y} = 0.21606 - 0.00038x_6$

If the variable $x_1, x_2, x_3, x_4, x_5, x_6, x_8$ considered constant, the influence of the variable x_7 is on the residual martingale survival rate of cervical cancer patients is;

$$\hat{y} = 0.21606 - 0.0004x_7 - 0.0004(x_7)_+$$

$$= \begin{cases} 0.21606 - 0.0004x_7; & x_7 < 0 \\ 0.21606 - 0.0008x_7 & x_7 \geq 0 \end{cases}$$

From this model it can be interpreted that if patients who do not experience complications have equation $\hat{y} = 0.21606 - 0.0004x_7$. Furthermore, if the patient has complications, they have equation $\hat{y} = 0.21606 - 0.0008x_7$

If the variable $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ considered constant then the influence of the variable x_8 to martingale residual survival rate of cervical cancer patients *y* is;

$$\hat{y} = 0.21606 - 0.000575x_8$$

$$= \begin{cases} 0.2160 + 0.000575x_8; & x_8 < 1 \\ 0.21606 + 0.000575x_8 & x_8 \geq 1 \end{cases}$$

From this model it can be interpreted that if patients undergoing chemotherapy and patients who are not undergoing chemotherapy have equation $\hat{y} = 0.21606 + 0.0008x_8$.

Conclusions

Based on the results of the research that has been done, it is found that the survival probability of cervical cancer patients in the time range from day 0 to day 150 is 0.88. This value shows the probability of survival of cervical cancer patients is still high. Meanwhile, the curve on the 150th day of survival probability of cervical cancer patients tends to be constant. The Spline model developed in this study is able to model cervical cancer patient data well as evidenced by a small MSE value.

Acknowledgements

This research was funded by the Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Indonesia, grant number: 161/E5/PG.02.00.PT/2022, in the Higher Education Excellence Basic Research scheme.

References

[1] K Dauda, WB Yahya and A Banjoko. Survival analysis with multivariate adaptive regression using cox-snell residual. *Anale Seria Informatică* 2015; **13**, 25-41.

[2] G Yaprak, D Tataroglu, B Dogan and M Pekyurek. Prognostic factors for survival in patients with gastric cancer: Single-centre experience. *N. Clin. Istanbul* 2020; **7**, 146-52.

[3] J Faradmal, A Talebi, A Rezaianzadeh and H Mahjub. Survival analysis of breast cancer patients using cox and frailty models. *J. Res. Health Sci.* 2012; **12**, 127-30.

[4] MA Pourhoseingholi, E Hajizadeh, DB Moghimi, A Safaee, A Abadi and MR Zali. Comparing cox regression and parametric models for survival of patients with gastric carcinoma. *Asian Pac. J. Canc. Prev.* 2007; **8**, 412-6.

[5] B Moghimi-Dehkordi, A Safaee, MA Pourhoseingholi, R Fatemi, Z Tabeie and MR Zali. Statistical comparison of survival models for analysis of cancer data. *Asian Pac. J. Canc. Prev.* 2008; **9**, 417-20.

[6] A Perperoglou, A Keramopoullos and HC Houwelingen. Approaches in modelling long-term survival: An application to breast cancer. *Stat. Med.* 2007; **26**, 2666-85.

[7] I Pelagia. 2016, Variable selection of fixed effects and frailties for cox proportional hazard frailty models and competing risks frailty models. Ph. D. Dissertation. University of Manchester, Manchester, England.

[8] P Naseri, AR Baghestani, N Momenyan and ME Akbari. Application of a mixture cure fraction model based on the generalized modified weibull distribution for analyzing survival of patients with breast cancer. *Int. J. Canc. Manag.* 2018; **11**, e62863.

[9] A Kavkler, D Daniela-Emanuela, AG Babucea, I Bicanic, D Tevdovski, K Tosevska-Trpcevska and D Boršič. Cox regression models for unemployment duration in Romania, Austria, Slovenia, Croatia, and Macedonia. *Rom. J. Econ. Forecast.* 2009; **10**, 81-104.
[10] SM Grybach, LZ Polishchuk and VF Chekhun. Analysis of the survival of patients with breast cancer depending on age, molecular subtype of tumor and metabolic syndrome. *Exp. Oncol.* 2018; **40**, 243-8.
[11] SC Robles and E Galanis. Breast cancer in latin america and the caribbean. *Revista Panamericana Salud Publica* 2002; **11**, 178-85.
[12] RO Ferraz and DC Moreira-Filho. Survival analysis of women with breast cancer: Competitive risk models. *Cien Saude Colet* 2017; **22**, 3743-54.
[13] M Liu, L Li, W Yu, J Chen, W Xiong, S Chen and L Yu. Marriage is a dependent risk factor for mortality of colon adenocarcinoma without a time-varying effect. *Oncotarget* 2017; **8**, 20056-66.
[14] JY Charati, G Janbabaei, N Alipour, S Mohammadi, SG Gholiabad and A Fendereski. Survival prediction of gastric cancer patients by artificial neural network model. *Gastroenterol. Hepatol. Bed Bench* 2018; **11**,110-7.
[15] R Kelter. Bayesian survival analysis in STAN for improved measuring of uncertainty in parameter estimates. *Measurement: Inter. Res. and Persp.* 2020; **18**, 101-9.
[16] J Lu, L Cao, CH Zheng, P Li, JW Xie, JB Wang, JX Lin, QY Chen, M Lin and RH Tu. The preoperative frailty versus inflammation-based prognostic score: Which is better as an objective predictor for gastric cancer patients 80 years and older? *Ann. Surg. Oncol.* 2017; **24**, 754-62.
[17] Y Fujisawa, S Yoshikawa, A Minagawa, T Takenouchi, K Yokota, H Uchi, N Noma, Y Nakamura, J Asai, J Kato, S Fujiwara. Clinical and histopathological characteristics and survival analysis of 4594 Japanese patients with melanoma. *Cancer med*. 2019; **8**, 2146-56.
[18] DR Arifanti and R Hidayat. Statistical inference on nonparametric spline models and its applications. *J. Phys. Conf.* 2021; **2123**, 012023.
[19] JD Opsomer and D Ruppert. Fitting a bivariate additive model by local polynomial regression. *Ann. Stat.* 1997; **25**, 186-211.
[20] JM Maronge, Y Zhai, DP Wiens and Z Fang. Optimal designs for spline wavelet regression models. *J. Stat. Plann. Infer.* 2017; **184**, 94-104.
[21] A Antoniadis, J Bigot and T Sapatinas. Wavelet estimators in nonparametric regression: A comparative simulation study. *J. Stat. Software* 2001; **6**, 1-83.
[22] A Antoniadis and F Leblanc. Nonparametric wavelet regression for binary response. *Statistics* 2000; **34**, 183-213.
[23] R Hidayat, Ma'rufi and Yuliani. Integral estimator with kernel approach for estimating nonparametric regression functions. *J. Phys. Conf.* 2021; **2123**, 012022.
[24] IN Budiantara, V Ratnasari, M Ratna and I Zain. The combination of spline and kernel estimator for nonparametric regression and its properties. *Appl. Math. Sci.* 2015; **9**, 6083-94.
[25] V Ratnasari, IN Budiantara, M Ratna and I Zain. Estimation of nonparametric regression curve using mixed estimator of multivariable truncated spline and multivariable kernel. *Global J. Pure Appl. Math.* 2016; **12**, 5047-57.
[26] R Hidayat, IN Budiantara, BW Otok and V Ratnasari. The regression curve estimation by using mixed smoothing spline and kernel (MsS-K) model. *Comm. Stat. Theor. Meth.* 2021; **50**, 3942-53.
[27] IN Budiantara, V Ratnasari, M Ratna, W Wibowo, N Afifah, DP Rahmawati, MAD Octavanny. Modeling percentage of poor people in indonesia using kernel and fourier series mixed estimator in nonparametric regression. *Invest. Operacional.* 2019; **40**, 538-50.
[28] H Nurcahayani, IN Budiantara and I Zain. The curve estimation of combined truncated spline and fourier series estimators for multi-response nonparametric regression. *Mathematics* 2021; **9**, 1141.
[29] R Hidayat, IN Budiantara, BW Otok and V Ratnasari. An extended model of penalized spline with the addition of kernel functions in nonparametric regression model. *Appl. Math. Inform. Sci.* 2019; **13**, 18.