

Evaluation of the Effect of Prognostic Variables on the Survival Analysis of Prostate Cancer

Shafiq Hossain Sourav, Rownak Jahan Tamanna

Institute of Statistical Research and Training, University of Dhaka, Dhaka, Bangladesh,
e-mail: msourav@isrt.ac.bd, rjahan@isrt.ac.bd

SUMMARY

Prostate cancer is a severe threat to human lives. Approximately 1 in 7 men will be diagnosed with prostate cancer throughout their lifetimes, and 1 in 39 men will die from prostate cancer. There are many factors which increase or decrease the survival time of prostate cancer patients. Data is used here from a randomised clinical trial for the choice of treatment for prostate cancer patients in stages 3 and 4. This study is done to identify probable variables that influence the survival time of patients only for these two stages. The AFT and the Cox-PH models determine how variables affect prostate cancer patients' survival time.

Key words: Cox-PH, AFT, Log-rank

1. Introduction

Cancer, characterised by uncontrolled cell growth, is the world's leading cause of death. There are about 100 forms of cancer, each categorised by the original afflicted cell type. Prostate cancer is a male reproductive system malignancy. It affects around 1 in every 7 men and kills approximately 1 in every 39 men (see Shahi and Ataharul Islam, 2016). Several risk factors for prostate cancer have been found, but how these risk factors induce a prostate cell to become malignant is unknown. Genetic material alterations cause approximately 5% to 10% of prostate tumours. Age, family history, fatty diet, and race are all risk factors for prostate cancer (see Moul et al., 1996). Prostate cancer poses a significant threat to human life and must be thoroughly researched. Researchers are currently investigating the cancer's aetiology. However, there are no known means to eradicate this

disease. We understand that some prognostic variables accelerate or retard prostate cancer patients' survival time. An exhaustive model-based survival study will allow us to determine which variables have the most influence on the survival time of prostate cancer patients and the size of that impact. We did not make any comparisons between them. Our goal was to obtain two interpretations, one regarding hazards (risk of death) and the other regarding survival time. Because of this, we fitted two models: the Cox Proportional Hazards (Cox-PH) model and the Accelerated Failure Time (AFT) model. However, this study will aid the following generation of students in learning how to fit a survival model, check their diagnostics, and interpret those results.

2. Background of the Study

Prostate cancer is the second leading cause of cancer death in men in the US (see Jemal et al., 2009). Localised prostate cancer is curable with surgery or radiation. However, recurrence occurs in 20–30% of individuals. Approximately 20,000 new cancer cases are detected yearly in Bangladesh (see Hussain, 2013). From 2008 to 2013, the incidence of prostate cancer was 2.3% (see Uddin et al., 2013). In 2008, prostate cancer caused 2% of cancer deaths in the Asia-Pacific region (see Baade et al., 2009).

Byar and Green (1980) present a thorough example utilising estrogen therapy data for prostate cancer. This study found treatment-covariate interactions. Young individuals with high-grade cancers should have been treated with estrogens, whereas older people with low-grade tumours were harmed. Lunn and McNeil (1995) gave a Cox proportional hazards regression model for the joint estimation of parameters in models for competing risks in survival analysis for cardiovascular disease. Between 1986 and 1992, 84% of prostate cancer patients identified at the Bombay Hospital Institute of Medical Sciences were in an advanced clinical stage (see Srinivas et al., 1995). However, just 15% of individuals were detected at an advanced stage when they first presented to a US hospital over that same period (see Esper et al., 1997).

In recent years, different studies have been carried out to determine the potential risk factors and their magnitude for the survival of patients with prostate cancer. A survey of Malaysian patients found that age greater than 75 years, Gleason scores greater than eight, and disease metastasis were predictive markers for prostate cancer survival (see Shah et al., 2021). Using

Polymerase Chain Reaction-Restriction Fragment Length Polymorphism, a case-control study conducted in Bangladesh in 2022 identified two alleles of a gene associated with an increased risk of prostate cancer (see Bellah et al., 2022).

3. Methodology

Survival Analysis is a set of statistical approaches used to examine the time-to-event data known as survival time. The survival function predicts whether a patient, device or other entity will survive beyond a specific period.

3.1. Non-Parametric Approach

Before assuming any study time distribution, we run some exploratory analyses in survival analysis. This gives us a rough scenario of the study phenomenon by producing some graphs of survival function, hazard function, and descriptive statistics like median survival time. These are non-parametric methods, as we do not assume any distribution of study time. A popular non-parametric method is the Kaplan–Meier estimate (see Kaplan and Meier, 1958). Another procedure to compare two or more survival distributions when the data are right censored is the log-rank test (see Mantel, 1966). It tests whether all groups have the same survival distribution (null hypothesis). Each group’s observed occurrences are compared using log-rank testing, which would be expected under the null hypothesis. This statistic has a chi-square distribution with degrees of freedom $k-1$ (k = number of groups) (see Bland and Altman, 2004). If O_i and E_i are the observed and expected number of events, respectively, then the log-rank test statistic is

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}. \quad (1)$$

3.2. Parametric Approach

In most situations, prognostic factors influence survival time. Non-parametric methods cannot assess the magnitude of covariate impacts on survival time; hence we require a multivariate approach. However, multivariate linear regression and logistic regression do not count censored data. Parametric

proportional hazard (PH) models and accelerated failure time (AFT) models can be used to examine survival data, since the probability distributions for survival time are limited. Firstly, we will fit the AFT model to analyse survival time data.

The model is theoretically stated as $S(t/x) = S_0(t=\eta(x))$ for a group of individuals with covariate $(X_1, X_2, \dots, X_k)$, where η is an constant acceleration factor corresponding to any value of $S(t)$, and $S_0(t)$ is the baseline survival function. The log-linear version of the AFT model for time is expressed as

$$\log T_i = \mu + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_p x_p + \sigma \varepsilon_i. \quad (2)$$

Here, σ is the scaling parameter, μ is the intercept, ε_i is the random variable, assumed to have a specific distribution, p is the number of coefficients, and i ranges over $1 \dots n$ where n is the number of observations. There will be a distribution for T correspondingly for each distribution of ε_i . Among the most common distributions listed in Table 1, we will look for a distribution which can better explain our time variable.

Table 1. Distribution of ε and T.

Distribution of ε	Distribution of T
LogGamma	Gamma
Normal	Lognormal
Logistic	Loglogistic
Extreme value(1 parameters)	Exponential

3.2.1. Goodness of Fit

As time can follow any distribution and the AFT model can have many possible distributions, all AFT models may be compared using statistical tests or criteria. With the Akaike information criterion (AIC), all models may be compared (see Akaike, 1974). Residual plots can also be used to assess model fit. Two methods exist. One method uses residuals from the AFT model to calculate Cox-PH residuals. One interesting figure compares Cox-Snell residuals to the unit exponential distribution. The Cox-Snell residual for the i^{th} individual at time t_i is defined as follows:

$$r_{ci} = \hat{H}(t_i|x_i) = -\log \hat{S}(t_i|x_i), \quad (3)$$

where $\hat{S}(t_i)$ is the estimated survival function on the fitted model, and t_i and x_i are the observed survival time and the vector of covariate values

respectively for individual i . The Cox–Snell residual can be used to check the goodness of fit for any parametric model (see Ansin, 2015). The plot of $\log(-\log S(r_{c_i}))$ versus $\log(r_{c_i})$ will be a straight line with unit slope through the origin, if the fitted model is correct (see Lawless, 2011). A plot of residual deviance vs. survival time or explanatory variables may identify model fit issues. Another procedure is based on the AFT model’s standardised residuals. At first, standardised residuals from all AFT models are computed. The q-q plot of standardised residuals is plotted against the theoretical quantiles of the standardised survival distribution. If the points plotted for any given AFT model lie on a line with a slope of one and an intercept of zero, it might be concluded that the AFT model matched the data set well. In our study, we used the second option.

3.3. Semi Parametric Approach

Only a few survival time distributions can be utilised in proportional hazard models. For the parametric proportional hazard model, baseline hazard must follow either Weibull distribution or exponential distribution, which is the main drawback of this model. One of the most popular approaches to dealing with these situations is the semi-parametric Cox-Proportional Hazards model (Cox-PH), proposed by Cox (1992). The Cox-PH model is given by-

$$h(t|X_i) = h_0(t) \exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik}) = h_0(t) \exp(\beta' X_i), \quad (4)$$

where $h_0(t)$ is the baseline hazard function for an individual with zero model variables, $X_i = (X_{i1}, X_{i2}, \dots, X_{ik})'$ contains the values of the vector of explanatory variables for a particular individual, $\beta = (\beta_1, \beta_2, \dots, \beta_k)$ is a vector of regression coefficients, k is the number of coefficients, and i ranges over $1, \dots, n$ where n is the number of observations.

3.3.1. Checking Proportional Hazard Assumption

The Cox-PH model assumes that the influence of a covariate does not change over time. The basic Cox model is inaccurate if this criterion is not satisfied. Schoenfeld residuals verify PH assumptions (see Abeysekera and Sooriyarachchi, 2009). The null hypothesis is that scaled Schoenfeld residuals have no Pearson product-moment correlation with time. Rejecting the null hypothesis violates the PH assumption.

3.3.2. Goodness of Fit

After fitting, the model's validity must be assessed. The goodness of fit of the Cox model can be checked by means of Cox–Snell residuals (see Ansin, 2015). The Cox–Snell residual for the i th individual with observed survival time t_i , $i = 1, \dots, n$, is defined as:

$$r_i = \exp(\hat{\beta}'x_i)\hat{H}_0(t_i) \quad (5)$$

where $\hat{\beta}'x_i = \hat{\beta}_1x_{1i} + \hat{\beta}_2x_{2i} + \dots + \hat{\beta}_px_{pi}$ is the value of the fitted component of the model for that individual, and $\hat{H}_0(t_i)$ is an estimate of the baseline hazard function at time t_i . After obtaining the estimated cumulative hazard function of the Cox–Snell residuals, $(\hat{H}_r(r))$, which is also known as the Nelson–Aalen cumulative hazard (see Bitew et al., 2021), the plot of $(\hat{H}_r(r))$ against r_i should line up through the origin and has a slope of one. Any deviation from this is a sign of a poor match.

3.3.3. Detecting Outliers

Outliers of the Cox-PH model can be detected by deviance residuals (see Jemal et al., 2009). The deviance residuals are normalised martingales transformed. If the fitted model is suitable, the deviation residuals are generally symmetrical about zero and have a mean of zero. And also, there will be no outliers. Potential outliers will have large deviation residuals.

Survival, Survminer, ggplot2, ggpubr (see Therneau and Lumley, 2015; Kassambara et al., 2017; Wickham et al., 2016; Kassambara, 2020) packages from the R language (R Core Team, 2000) are used to draw the visualisations for the nonparametric approach. The model fitting and assumption tests are done using the *stset* and *stphtest* functions (see StataCorp, 2005) on Stata software (StataCorp, 2007).

4. Data Description

This data is based on a randomised controlled clinical trial conducted to compare the various doses of estrogen to treat stage 3 and 4 patients with prostate cancer, by Byar and Green (1980). There is a total of 502 observations and 18 variables. After omitting missing values and events that occurred or were censored at time zero, there remain 459 observations, and of them, 341 are censored. Hence, 74% of observations are censored here. Our response variables are *dttime* and *status*, and the explanatory variables

are stage, rx, age, wt, pf, hx, sbp, dbp, ekg, sz, sg, hg and bm. Here we excluded variable ap because the stage of the patients is marked according to the serum prostatic acid (ap) variable. If ap is less than ten, it is stage 3 prostate cancer; if it is greater than ten, it is stage 4 prostate cancer. If we consider both variables, it will cause a correlation (see Harrell Jr, 2015). Table 3 gives descriptions of the response variables; the remainder of them, which are not used in our analysis, are excluded.

Here we will first make all the continuous variables into categories, as described in a paper by Kay (1986). They are coded as:

Table 2. Conversion of continuous variables to categorical variables

Variable	Value	Label	Variable	Value	Label
	<65	0		<60	0
age	65-75	1	dbp	60-90	1
	75>	2		90>	2
	<80	0		<10	0
wt	80-100	1	sbp	10-14	1
	100>	2		14>	2
	<12	0		<30	0
hg	12>	1	sz	30>	1
	<10	0			
sg	10>	1			

One variable, *sg*, is an index whose value lies between 7 and 15. It can be treated as a factor or also as continuous (see Bozdogan and Hamparsum, 2003). Hence, we re-coded this to treat *sg* as a factor.

5. Analysis and Results

5.1. Result from Non-parametric Methods

Kaplan–Meier curves and the log-rank test are made to understand the primary behaviour of the variables, that is, for comparing two or more survival distributions of the patients. We took the p-value of the log-rank test from the *ggsurvplot* function while drawing the Kaplan–Meier curve using the argument *pval = True*. All the graphs (13 variables), along with the p-values of the log-rank test, can be found in the Appendix 6 section. Only the significant variables (p-value <0.05) are listed in Table 4. Therefore, we can claim that our non-parametric technique indicates that only these variables can influence the patient’s survival time. However, to determine the effect of each variable, we must fit a model, which we shall do accordingly.

Table 3. Variable description

Name	Labels	Levels
stage	Stage	3 4
rx	Treatment	placebo 0.2 mg estrogen 1.0 mg estrogen 5.0 mg estrogen
dtime	Months of Follow-up	integer
status	The post-trial survival status of the patient	alive dead – prostatic cancer dead – heart or vascular dead – cerebrovascular dead – pulmonary embolus dead – other cancer dead – respiratory disease dead – other specific non-cancer dead – unspecified non-cancer dead – unknown cause
age	Age in Years	integer
wt	Weight Index = $\text{wt}(\text{kg}) - \text{ht}(\text{cm}) + 200$	integer
pf	How active the patient is	normal activity in bed <50% daytime in bed >50% daytime confined to bed
hx	History of Cardiovascular Disease	No Yes
sbp	Systolic Blood Pressure/10	continuous
dbp	Diastolic Blood Pressure/10	continuous
ekg	Electrocardiogram code	normal benign rhythmic disturb & electrolyte changes heart block or conduction defects heart strain old MI recent MI
hg	Serum Haemoglobin (g/100ml)	continuous
sz	Size of Primary Tumour (cm^2)	continuous
sg	Combined Index of Stage and Hist. Grade	continuous
bm	Bone Metastases	No Yes.

Table 4. Only the significant variables' p-values from the log-rank test are shown here

Variables	P-value	Variables	P-value
stage	<.0001	dbp	.018
rx	.013	hg	.0012
age	.004	sz	<.0001
wt	.012	bm	<.0001
pf	<.0001	sg	<.0001

5.2. Fitting AFT Model

To express the magnitude of the effect of covariates in a more accessible way in terms of the difference between survival time for different levels of covariates, we will now fit the AFT model to the data set.

5.2.1. Finding the Distribution of the Survival Time

Before fitting the AFT model, the dataset needs to be checked for which distribution the survival time follows. As discussed in Table 1, we will check the time with its corresponding random error as described in a book by Lawless (2011).

From Figure 1, we can see that the Weibull and the Log-normal plot are more linear. Hence, we can say that the time follows Weibull and log-normal distribution.

5.2.2. Comparing AFT Models Using AIC

So far, our findings have suggested that the survival time follows Weibull and Log-normal distribution. Now the Weibull AFT model and the Log-normal AFT model are compared using the statistical criterion AIC (the smaller AIC is, the better).

Table 5. Akaike Information Criterion in the AFT models

Name of distribution	Number of parameter	AIC
Weibull	2	553.30
Log Normal	2	565.24

From Table 5, the Weibull AFT model appears to be an appropriate AFT model according to AIC values, compared with other AFT models.

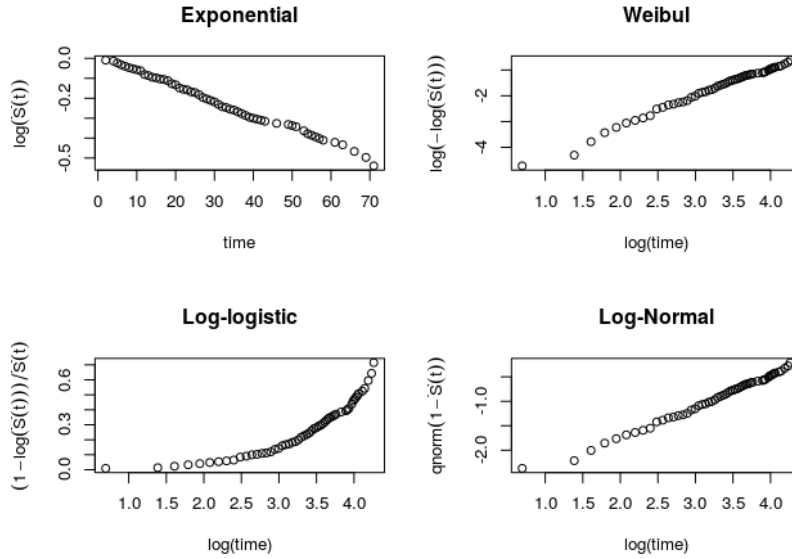


Figure 1. The distribution of time

5.2.3. Comparing AFT Models Using Q-Q Plot of Standardized Residuals

The AFT model's goodness of fit can be checked using standardized residual plots. Figure 2 shows that the plotted points for the Weibull AFT model seem to lie on a line with a unit slope and almost zero intercepts compared with other plots. Therefore the Weibull AFT model is the best-fitted AFT model for our data.

5.2.4. Fitting the Weibull AFT Model

As the Weibull distribution can explain the data well, we fitted the model with all thirteen variables. However, we later discovered that stage, wt, hx, sbp, DBP, and EKG do not significantly impact our survival time (the p-values are higher than 0.05). We, therefore, excluded those from our model and refitted the model with the remaining variables, which all significantly impact our survival time.

From Table 6, we can conclude that the covariates age, rx, pf, hg, bm, sg and sz are statistically significant at a 5% level of significance. Survival times decreased by 0.01 months as age increased by one year. Survival time for the patients who received 0.2 mg estrogen treatment is 6% lower than for

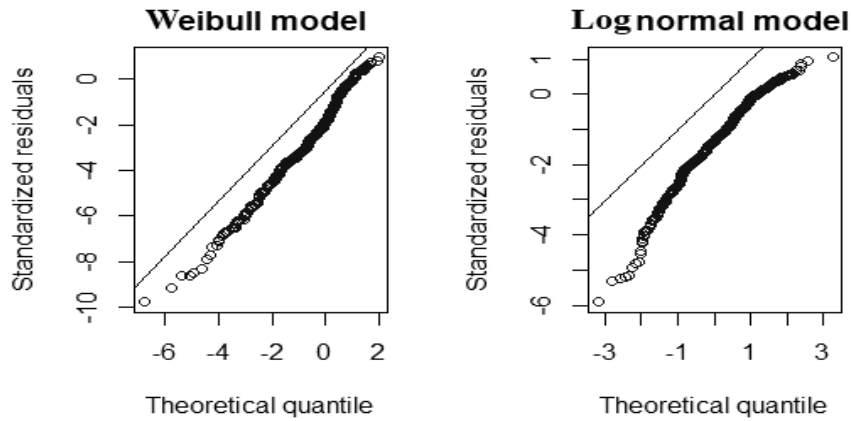


Figure 2. The q-q plot of the standardized residuals for AFT models.

Table 6. Results obtained by fitting Weibull AFT model

Variable	Category	e^{β}	SE of β	Z-value	P-value	95% Conf. Interval	
Intercept		32.75937	0.85666	4.07	0.000	1.81016	5.16822
age		0.98726	0.00875	2.18	0.029	0.00192	0.03623
rx	placebo						
	0.2 mg estrogen	0.94542	0.16206	-0.35	0.729	-0.37376	0.26151
	1.0 mg estrogen	1.94503	0.19866	3.35	0.001	0.27592	1.05463
	5.0 mg estrogen	1.90895	0.19856	3.26	0.001	0.25738	1.03573
pf	normal activity						
	in bed <50% daytime	1.34628	0.25115	1.18	0.236	-0.19489	0.78958
	in bed >50% daytime	0.42592	0.27401	-3.11	0.002	-1.39055	-0.31646
	confined to bed	0.36895	0.53	-1.88	0.06	-2.03587	0.04168
hg		1.07994	0.03551	2.17	0.03	0.00731	0.1465
sz		0.97225	0.00517	-5.44	0	-0.03828	-0.01801
bm	No	1					
	Yes	0.70116	0.16036	-2.21	0.027	-0.66933	-0.04071
sg	<10	1					
	10>	0.25241	0.22036	-6.25	0	-1.80861	-0.94481

those who received a placebo. Survival time for the patients who received 1 mg estrogen treatment is 94% higher than for those who received a placebo. Survival time for the patients who received 5 mg treatment is 90% higher than for those who received a placebo. Survival time for the patients who were told to take bed rest for less than 50% of daytime is 34% higher than

for those advised to do regular activity. Survival time for the patients who were told to take bed rest for greater than 50% of daytime is 58% lower than for those advised to do regular activity. Survival time for patients told to be confined to bed is 64% lower than for those advised to do regular activity. Survival times increased by 0.07 months with one unit increase of haemoglobin. Survival times decreased by 0.03 months with one unit increase in tumour size. Survival time for patients with bone metastasis is 30% lower than for patients without bone metastasis. Survival time for patients with a tumour grade higher than ten is 75% lower than for those with a grade of less than ten.

5.3. Fitting the Cox-PH Model

Firstly, like the Weibull AFT model, we fitted the Cox-PH model with all the variables. Fortunately, we found that the same seven variables as obtained in the Weibull AFT model significantly impact the risk of death. The remainder are excluded from our analysis as they have much higher p-values. For this reason, we will check the relevant assumption and interpretation of those specific variables. We will now fit the data set to the Cox-PH model to obtain an interpretation in terms of the hazard ratio. Before that, we also need to check the assumptions and model adequacy.

5.3.1. Checking Proportionality Assumptions for Covariates of Final Model

Table 7. Proportionality assumption test

Variable	Category	ρ	χ^2	p
age	age	-0.1333	2.4846	0.115
rx	0.2 mg estrogen	0.0675	0.54864	0.459
	1.0 mg estrogen	-0.0266	0.08204	0.775
	5.0 mg estrogen	-0.0551	0.38969	0.532
pf	confined to bed	0.0845	0.85606	0.355
	in bed <50% daytime	0.1198	1.66917	0.196
	in bed >50% daytime	0.0841	0.88425	0.347
hg		0.1304	1.93682	0.164
sz		0.0213	0.05552	0.814
bm	Yes	0.0686	0.53368	0.465
sg	Yes	-0.0032	0.00121	0.972

Table 7 shows that the p-values for all the covariates are more significant than 0.05, which suggests that the PH assumption is satisfied for all the covariates at a 5% level of significance. Now the proportional hazard assumptions are checked by the graphical procedure. Under the proportional hazards assumption, estimates of the parameters do not vary much over time.

5.3.2. Assessing the Goodness of Fit for the Final Model

The goodness of fit can be assessed using a Cox–Snell residual plot. A plot (Figure 3) of the Cox–Snell residuals (r_i) against the cumulative hazard (Nelson–Aalen cumulative hazard) of the Cox-PH model ($\hat{H}_r(r)$) is presented here.

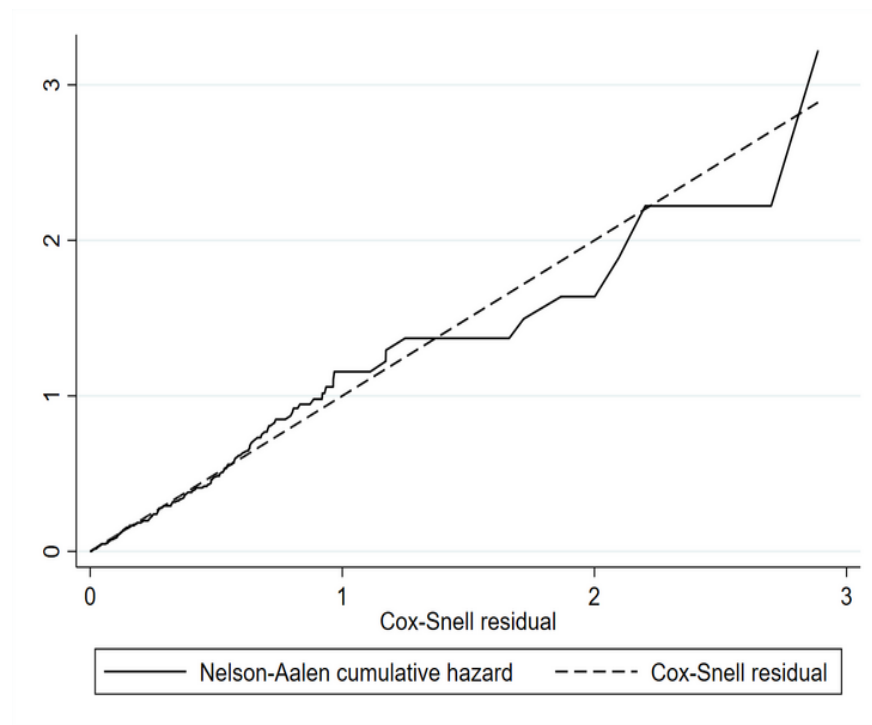


Figure 3. Goodness of fit of the final Cox-PH model

From Figure 3, there is little evidence of a systematic deviation of the curve from the straight line with a unit slope, indicating that the model is moderately adequate.

5.3.3. Detecting Outliers of the Final Model

Deviance residuals are used to examine a model for outliers. The plot of the deviance residuals against the estimated time (months) is shown in Figure 4. It is observed from Figure 4 that the deviance residuals seem to be distributed about zero. There are not-so-high or not-so-low deviance residuals, which suggests that these observations may not have any outliers.

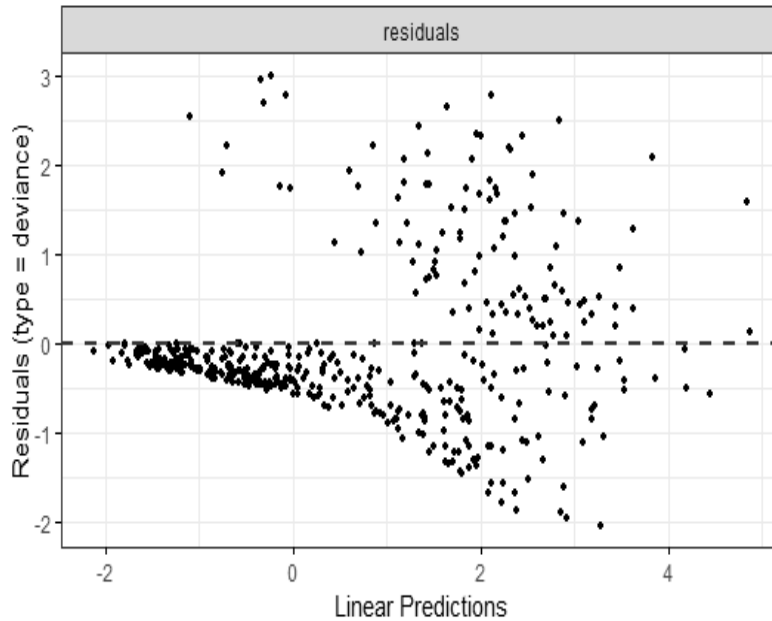


Figure 4. Deviance residuals against estimated time(months)

5.3.4. Findings of the Final Cox-PH Model

Having all the assumptions satisfied, we obtain the model to be fitted.

From the table, we can conclude that age, rx, pf, hg, bm, sg and sz are statistically significant at a 5% level of significance. The hazard of death increases by 0.03 months with one year of age increase. The risk of death increased by 9% and decreased by 61% and 59%, respectively, for the patients who took 0.2 mg, 1 mg and 5 mg estrogen treatment, compared with those who took the placebo. Again, the risk of death decreased by 36% and increased by 3.3 and 3.94 times, respectively, for those who were told to take

Table 8. Results obtained after fitting the final Cox-PH model

Variable	Category	Hazard Ratio	SE	z	P-value	95% Conf.Interval	
age		1.03125	0.01215	-2.16	0.031	0.94993	0.99756
rx	placebo	1					
	0.2 mg estrogen	1.09839	0.25660	0.4	0.688	0.69487	1.73623
	1.0 mg estrogen	0.39645	0.11251	-3.26	0.001	0.22732	0.69143
	5.0 mg estrogen	0.41815	0.12094	-3.01	0.003	0.23722	0.73709
pf	normal activity	1					
	in bed <50% daytime	0.64810	0.23454	-1.2	0.231	0.31886	1.31727
	in bed >50% daytime	3.31923	1.31893	3.02	0.003	1.52336	7.23220
	confined to bed	3.94029	3.02244	1.79	0.074	0.87620	17.7195
hg		0.89841	0.04573	-2.1	0.035	0.81311	0.99266
sz		1.04041	0.00759	5.43	0	1.02564	1.05539
bm	No	1					
	Yes	1.63173	0.38048	2.1	0.036	1.03316	2.57710
sg	< 10	1					
	10 >	7.05607	2.04905	6.73	0	3.99372	12.46659

bed rest for less than 50% of daytime, take bed rest for greater than 50% of daytime, and remain confined to bed, compared with those who were told to do regular activity. The hazard decreased by 11% and increased by 4%, respectively, for one unit increase of haemoglobin and tumour size. Those with bone metastasis have a hazard rate of death 63% higher than those who do not have it. Moreover, those whose tumour and histology grade is greater than 10 have a 7.05 times higher hazard rate of death than those with a grade less than 10.

6. Discussion

We begin with our graphical nonparametric Kaplan–Meier method and the log-rank test for the assessment of survival. We notice in graphical methods and log-rank tests that the survival distribution changes for stages, treatments, age, weight, performance, diastolic blood pressure, haemoglobin level, tumour size, bone metastasis, and tumour histology grade index. We also observe that cardiovascular disease does not affect survival time.

From the results of the AFT model, we found that survival time is respectively 6% lower and 94% and 90% higher for patients receiving 0.2 mg, 1 mg and 5 mg, compared with placebo patients. An individual's chance of

surviving is more significant when they engage in more physical activity and lower when they do not. More haemoglobin in the blood increases survival time. Tumour size increases the survival time, and those with bone metastasis and better histology grade have lower survival times.

Results obtained from fitting the Cox-PH model tell us that if we consider the 5% level of significance, patients taking 0.2 mg estrogen treatment have a high hazard rate. Patients with less physical movement also have a higher death hazard ratio. As tumour grade increases, so does the risk of death, and similarly with the tumour's size or the existence of bone metastases.

According to our findings, these same prognostic factors are significant at the same significance level. Consequently, the semi-parametric Cox-PH model and the parametric AFT model will provide identical results.

The data used in this article is from 1980. Sometimes data is not updated sufficiently for analysis. The data obtained from a small part of the world makes this study less reliable in assessing the general situation of prostate cancer patients globally. This research focuses on a few key factors. Other factors that impact disease-free survival time of prostate cancer patients include race, genotype, fatty foods, regular activity, and new treatments, which – as can be found from a literature review – are not utilised. This type of data might be studied further and modelled.

REFERENCES

- Abeysekera W., Sooriyarachchi M. (2009): Use of Schoenfeld's global test to test the proportional hazards assumption in the Cox proportional hazards model: an application to a clinical study. *Journal of the National Science Foundation of Sri Lanka* 37(1): 41–51.
- Akaike H. (1974): A new look at the statistical model identification. *IEEE transactions on automatic control* 19(6): 716–723.
- Ansin E. (2015): An evaluation of the Cox–Snell residuals.
- Baade P.D., Youlden D.R., Krnjacki L.J. (2009): International epidemiology of prostate cancer: geographical distribution and secular trends. *Molecular Nutrition and Food Research* 53(2): 171–184.
- Bellah S.F., Salam M.A., Billah S.S., Karim M.R. (2022): Genomewide association study of cyp3a4 and cyp3a5 genes in relation to the risk of prostate cancer among Bangladeshi population.
- Bitew Z.W., Ayele E.G., Worku T., Alebel A., Alemu A., Worku F., Yesuf A. (2021): Determinants of mortality among under-five children admitted with severe acute malnutrition in Addis Ababa, Ethiopia. *Nutrition Journal* 20(1): 1–15.
- Bland J.M., Altman D.G. (2004): The logrank test. *British Medical Journal (BMJ)* 328(7447): 1073.

- Bozdogan and Hamparsum (2003). Statistical data mining and knowledge discovery. CRC Press p.186
- Byar D., Green S. (1980): The choice of treatment for cancer patients based on covariate information. *Bulletin du Cancer* 67(4): 477–490.
- Cox D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2), 187–220.
- Esper P., Mo F., Chodak G., Sinner M., Cella D., Pienta K.J. (1997): Measuring quality of life in men with prostate cancer using the functional assessment of cancer therapy-prostate instrument. *Urology* 50(6): 920–928.
- Harrell Jr F.E. (2015): Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer.
- Hussain S.M.A. (2013): Comprehensive update on cancer scenario of Bangladesh. *South Asian Journal of Cancer* 2(4): 279.
- Jemal A., Siegel R., Ward E., Hao Y., Xu J., Thun M.J. (2009): Cancer statistics. *CA: A cancer journal for clinicians* 59(4): 225–249.
- Kaplan E.L., Meier P. (1958): Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282): 457–481.
- Kassambara A., Kassambara M.A. (2020): Package ‘ggpubr’. R package version 0.4.0.
- Kassambara A., Kosinski M., Biecek P., Fabian S. (2017): Package ‘survminer’. Drawing Survival Curves using ‘ggplot2’. R package version 0.4.9.
- Kay R. (1986): Treatment effects in competing-risks analysis of prostate cancer data. *Biometrics* 42(1): 203–211.
- Lawless J.F., 2011: Statistical models and methods for lifetime data. John Wiley & Sons.
- Lunn M., McNeil D. (1995): Applying Cox regression to competing risks. *Biometrics* 51(2): 524–532.
- Mantel N. (1966): Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 50: 163–170.
- Moul J.W., Douglas T.H., McCarthy W.F., McLeod D. G. (1996): Black race is an adverse prognostic factor for prostate cancer recurrence following radical prostatectomy in an equal access health care setting. *The Journal of Urology* 155(5): 1667–1673.
- Team R. C. (2000). R language definition. Vienna, Austria: R Foundation for Statistical Computing.
- Shah S.A., Ismail N., Taib S., Mat S.N., Safian N. (2021): Survival analysis and prognostic factors for prostate cancer patients at Universiti Kebangsaan Malaysia medical centre, Kuala Lumpur. *Sains Malaysiana* 50(5): 1367–1379.
- Shahi M.S.J.R., Ataharul Islam M. (2016): Survival analysis on phase iii clinical trial of estrogen on stage 3 and 4 prostate cancer patients. *International Conference on Analysis of Repeated Measures Data*, Dhaka, Bangladesh (1): 1.
- Srinivas V., Mehta H., Amin A., Choudary R., Gadgil N., Ravishanker D., Phadke A. (1995): Carcinoma of the prostate state at initial presentation. *International Urology and Nephrology* 27(4): 419–422.

- StataCorp L. P. (2007): Stata data analysis and statistical software. Special Edition Release 10, 733.
- StataCorp L.P. (2005): Stata base reference manual. College Station: StataCorp LLC.
- Therneau T.M., Lumley, T., (2015): Package ‘survival’. R Top Doc, 128(10), pp.28–33. R package version 3.3-1.
- Uddin A.K., Khan Z.J., Johirul Islam A. (2013): Cancer care scenario in Bangladesh. South Asian Journal of Cancer 2(2): 102.
- Wickham H., Chang W., Wickham M.H. (2016): Package ‘ggplot2’. Create elegant data visualisations using the grammar of graphics. Version, 2(1), 1-189. R package version 3.4.0.

APPENDIX

Kaplan–Meier Survival Curve for All Variables

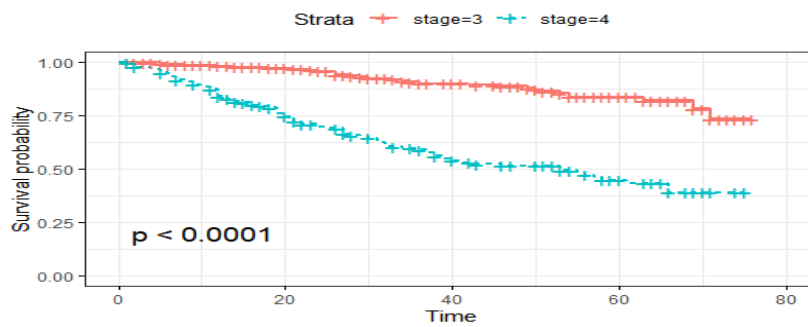


Figure 5. Kaplan–Meier survival curve for stage of prostate cancer

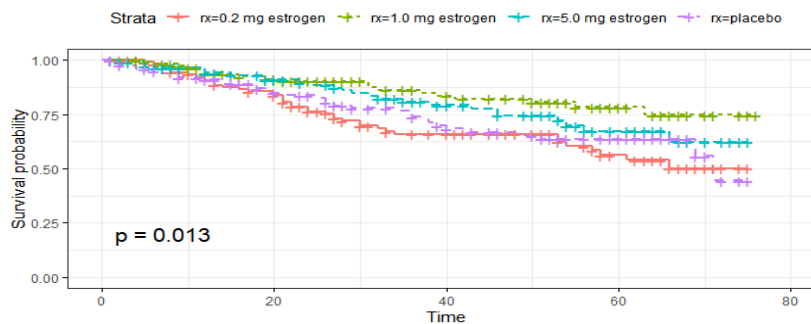


Figure 6. Kaplan–Meier survival curve for hormone treatment

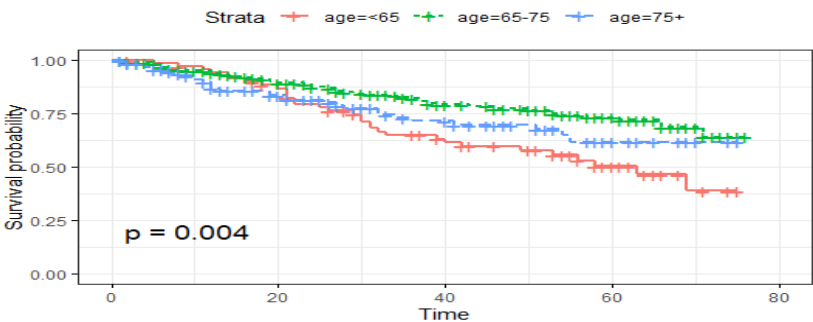


Figure 7. Kaplan–Meier survival curve for age

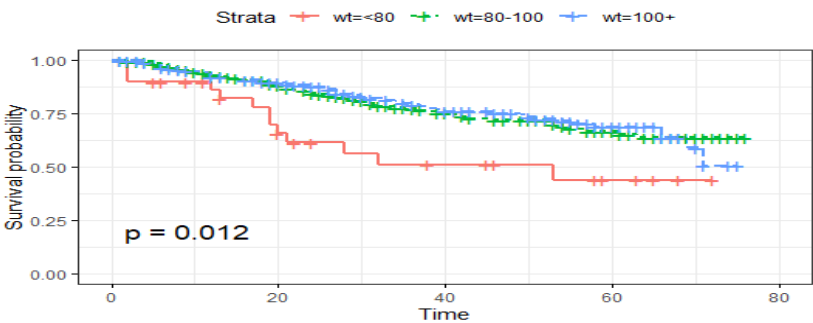


Figure 8. Kaplan–Meier survival curve for weight

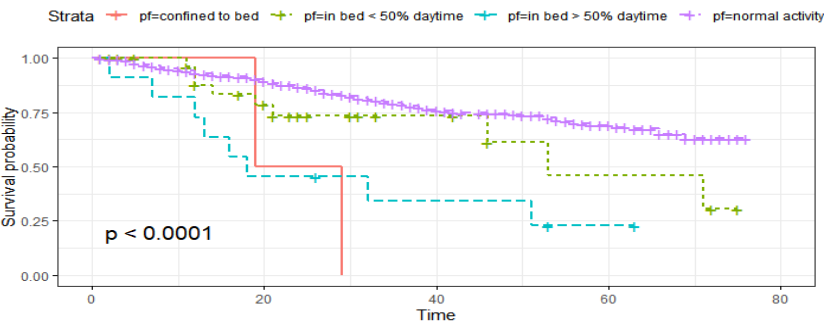


Figure 9. Kaplan–Meier survival curve for patient activity

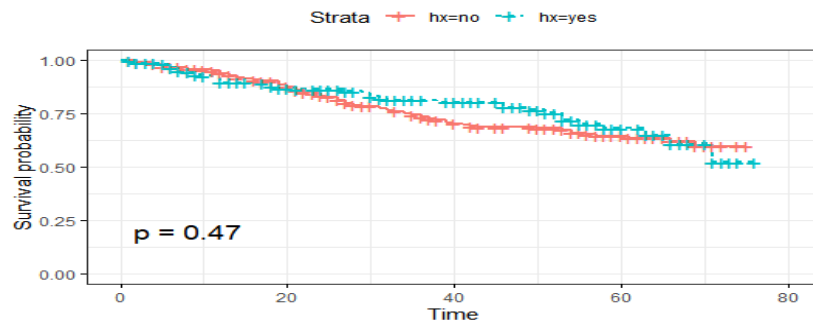


Figure 10. Kaplan–Meier survival curve for cardiovascular disease

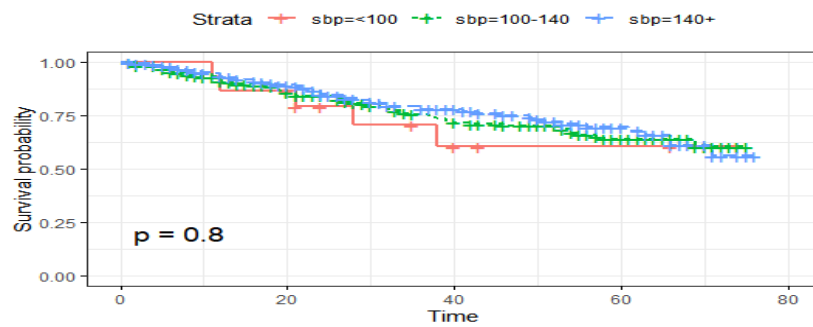


Figure 11. Kaplan–Meier survival curve for systolic blood pressure

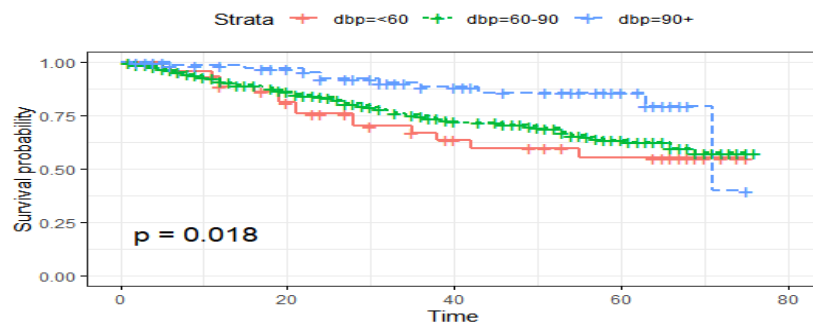


Figure 12. Kaplan–Meier survival curve for diastolic blood pressure

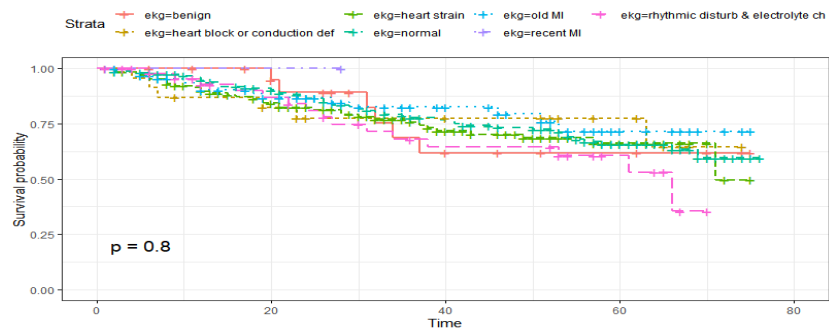


Figure 13. Kaplan–Meier survival curve for ECG code

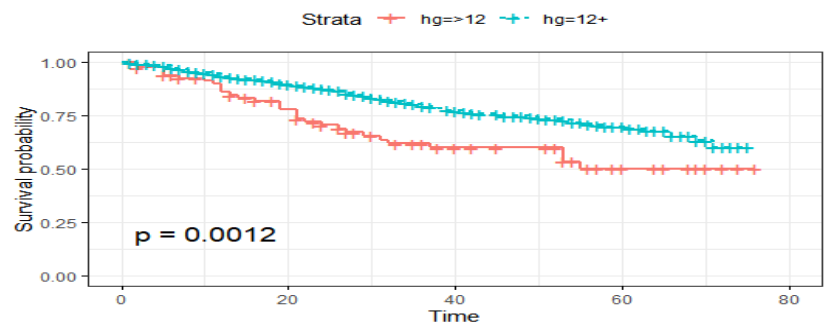


Figure 14. Kaplan–Meier survival curve for haemoglobin level

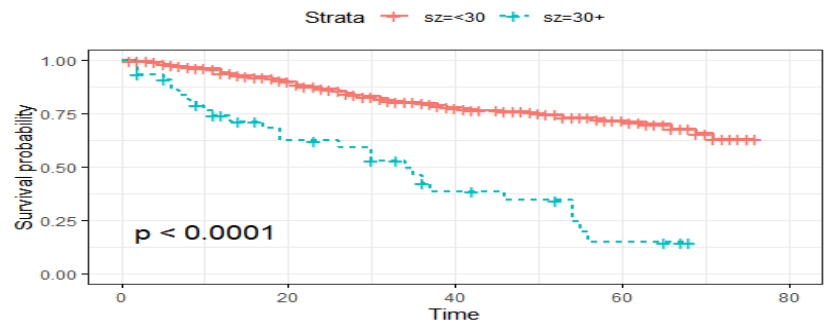


Figure 15. Kaplan–Meier survival curve for size of tumour

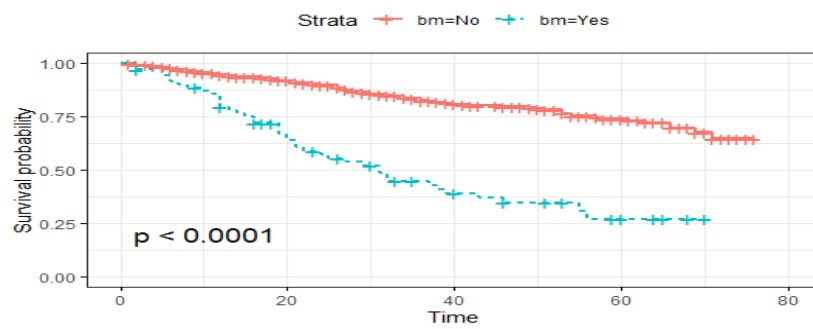


Figure 16. Kaplan–Meier survival curve for presence of bone metastasis

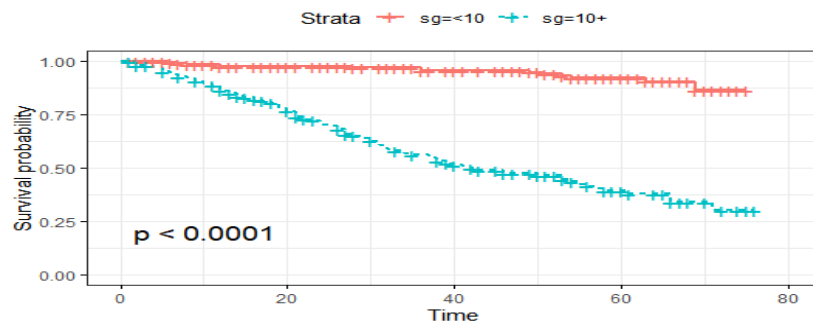


Figure 17. Kaplan–Meier survival curve for index of tumour stage