An investigation of the effect of the covariates on survival time of Breast Cancer patients in Saudi Arabia.

Refah Alotaibi

Mathematical Science Department, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia.
E-mail: rmalotaibi@pnu.edu.sa

Abstract: Breast cancer is the most common type of cancer in Saudi women, according to the Saudi Cancer Registry in King Faisal Specialist Hospital and Research Centre (KFSH and RC). Around 930 new cases is diagnosed per year, which means that 19.5% of women in Saudi Arabia (Saudi and non-Saudi). The primary objective of this work is to investigate the effect of the covariates on survival time of Breast cancer patients in Saudi Arabia in order to estimate the survival probability of the different patients with different covariate combinations. We fit the Cox proportional hazards model and predict the lifetime of any patient given her covariates. The survival data for this project will consist of a study of 8312 (8172) females and about 140 males (1.68%) patients with advanced breast cancer with thirteen covariates collected for 9-years (2004 to 2013).

[Refah Alotaibi. **An investigation of the effect of the covariates on survival time of Breast Cancer patients in Saudi Arabia.** *Cancer Biology* 2017;7(2):48-54]. ISSN: 2150-1041 (print); ISSN: 2150-105X (online). http://www.cancerbio.net. 7. doi:10.7537/marscbj070217.07.

Keywords: Breast Cancer, model, Kaplan-Meier, Cox Survival, hazard.

1. Introduction

Breast cancer is one of the most dangerous diseases and is the most frequently occurring cancer among women. According to the Saudi Cancer Registry in King Faisal Specialist Hospital and Research Center (KFSH and RC), Breast cancer is the most common type of cancer in Saudi women whereas around 930 new cases is diagnosed per year, which means that 19.5% of women in Saudi Arabia (Saudi and non-Saudi). In year 2010, there were 1473 female breast cancer cases. Breast cancer ranked first among females accounting for 27.4% of all newly diagnosed female cancers (5,378) in year 2010. The Age-Standardized Rate (ASR) was 24.9/100,000 for female population. The five regions with the highest ASR were Eastern region at 39.5/100,000, Qassim region at 32.8/100,000, Riyadh Region at 30.6/100,000, Makkah Region at 24.2/100,000 and Madinah Region at 21.3/100,000. The median age at diagnosis was 49 years (Range 21-120 years). In the world for example, the American Cancer Society in the year 2013 estimated that, about 232,340 new cases of aggressive breast cancer would be diagnosed; and about 39, 620 women would die due to this deadly disease (American Cancer Society., 2013).

Several studies have presented that, the Breast cancer is considered as the main cause of death in the Western societies (Tarone, 2006). One of the top malignancies among Saudi females is breast cancer, with a percentage of 21.8% (Registry, 2007) and it was the ninth leading death cause in females in Saudi Arabia in 2010. (Mokdad et al., 2014); (Lozano et al.,2012), Moreover, (Al-Qahtani, 2007) points out that the Breast Cancer is the second most common of the malignancy in Saudi women. (Ibrahim et al., 2008) concludes that over the coming decades in Saudi Arabia the rate of breast cancer will increase because of the increase in population and aging. Several studies have documented that breast cancer in young Saudi females is an important problem. As noted by the 2002 annual report of Saudi National Cancer Registry, breast cancers that developed before the age of 40 comprise 26.4% of all female breast cancers comparing to 6.5% in the USA. Several studies have revealed that Breast cancer in young patients (40 years old) is often related with a poorer prognosis and more aggressive, with higher mortality and recurrence rates compared with older women Elkum et al. (2007); (Zabicki et al., 2006); Han et al. (2004); Colleoni et al. (2002); Robson et al. (1998); Khan_r et al. (2006); Chia et al. (2004); El Saghir et al. (2006); Chung et al. (1996); Nixon et al. (1994).

Saudi Cancer Registry (SCR) is considered as a national cancer registry, followed to Saudi Health Council. It was established in 1402 Ah (1992) under authority of ministry of health and the main office was at the King Faisal specialist hospital and Research Centre in Riyadh and was moved to Saudi Health Council in 1435 Ah. SCR working to collect all data related to cancer registration from all the thirteen administrative regions in the king-dom which include: Riyadh, Estern region, Makkah, Madinah, Qassim, Hail, Jouf, Northen region, Tabouk, Najran, Baha, Asir, and Jezan. Therefore, it covers all the population in the country. The SCR Main Office indirectly supervises the regional offices and responsible for ensuring the accuracy and quality of data collected in all regions. Quality control processes include verification of site, morphology, and staging information as well as case linkage (tumor and patient), and consolidation of data.

These data were underwent a serious of steps to ensure its quality. Usually, in the original sites where regional and hospital cancer registries are located, data were abstracted from patients medical records, whom already classified as cancer cases based on their clinical diagnosis, histopathological, and/or radiological diagnosis. Other data were also collected as related to personal identification (name, ID Number, sex, age), demographic information (address, telephone number, nationality), and tumor details (diagnosis date, primary site, histology, behavior, grade, stage, basis of diagnosis). The primary site and histology of cancers are also identified and coded according to the International Classification of Diseases for Oncology (ICD-O-3), and all data were entered in the computer using a program software called Can Reg4 (IA CR) (Cancer Incidence Report, 2010).

2. Review of some concepts in survival analysis

The observation time t_i could either be the time from inclusion in the study until patient *i* dies from the disease of interest (indicated by $\delta_i = 1$) i.e time to failure or t_i can be the censoring time if patient *i* is still alive at the end of the study or if he or she drops out of the study or dies from another cause during the follow-up (indicated by δ_i=0). Common problems when observing patients in survival analysis include patients leaving or quitting before the time-frame of the study had elasped, patients dying due to causes different from the breast cancer, patients moving too far away to continue monitoring et.c. Hence, the true value of t_i is not always available. These events cause the value of t_i to be censored since the event time is definitely larger than the time between the beginning and the censoring events. We have right censoring when subjects are still alive when the study ends, they have lost follow-up or if the date of the event is after some time. We could also have left censoring if we never knew the exact time that the patient had the cancer or if the subject's lifetime is known to be less than a certain duration. Censoring has enabled researchers to analysis incomplete data. The standard assumption, also made here, is that the failure and censoring mechanisms are independent. Additionally, we will assume for simplicity that the observation time is continuous and no ties occur. Let the survival time t_i of an individual *i* be a realisation of a non-negative random variable T_i with probability density function $f_i(t)$ and cumulative distribution function (cdf) $F_i(t)$. Then, the lifetime distribution function of T_i, $F_i(t)$, is given by

$$F_i(t) = P(T_i < t).$$

The survival function $S_i(t)$ of an individual *i* can be defined as the probability that the individual survives longer than some specified time *t* where *t* ranges from *0* to ∞ (Lee and Wang, 2003). The survival function can be given as

$$S_i(t) = P(T_i \geq t) = 1 - F_i(t).$$

The Hazard function $\alpha_i(t)$ of an individual *i* can be expressed mathematically as

$$\alpha_i = \frac{f_{i(t)}}{S_i(t)}.$$

2.1 Kaplan-Meier Estimator

A basic task in the analysis of survival data is to estimate a survivor function. The two main non-parametric methods are the life-table and Kaplan Meier method. The Kaplan-Meier (KM) method is the most widely used, important and generally accepted estimator of the survivor function. It is also known as the product-limit estimator (Kaplan and Meier, 1958). The Kaplan Meier estimator of survival function at time t is given as follows:

$$\hat{S}(t) = \begin{cases} 1\, if\;\; t < t_i \\ \prod t_{i\leq}\, t \left[1 - \frac{d_i}{Y_i}\right]\, if\; t_i < t \end{cases} \quad (1)$$

where Y_i denotes the number of individuals who are at risk at time t_i (still alive and uncensored just before t_i) and d_i is the number of events at the i^{th} ordered time t_i.

Kaplan Meier can be used in determining the effect of the variables on the survival. The Kaplan Meier curves provide a graphical output which shows the plot of the percentage survival against time. Two major advantages of the Kaplan Meier curve are that it is very quick and easy to interpret. These make it possible for analyst to go through large amount of outcomes and get inferential behaviors in a short time.

2.2 Log-Rank Test Statistic

It will be interesting to know whether there are significant differences between groups in terms of survival. In some cases, the Kaplan Meier plots show that there are significant differences between the groups but it is important that the differences are backed up with a test. We need to find out if there are sufficient evidence to draw the conclusion that one group of people live longer than the other. The log-rank test is a formal hypothesis test used to compare survival curves using hypothesis tests. We use this test when we have two or more groups and we wish to test the null hypothesis that all groups have the same survival

rate. We recall the null (conservative) hypothesis H_0 as the case that the groups have the same lifetime distribution and an alternative hypothesis H_1 as the case that the groups have different lifetime distributions. If the value of the log-rank test is large then we reject the null hypothesis in favour of the alternative hypothesis.

Suppose we have r groups of individuals, with $r \geq 2$. We pool all the death times together to define intervals *[0, t_1), [t_1, t_2)*, etc. We have d_{bj} deaths in group *b* in interval *j* and n_{bj} individuals alive and uncensored from group *b* at the start of interval *j*. The log-rank test statistic is defined as:

$$\text{Log-rank test} = \sum_{b=1}^{r} \frac{(O_b - E_b)^2}{E_b}$$

where $O_b = \sum_{j=1}^{m} d_{bj}$ is the observed number of deaths in group *b* and $E_b = \sum_{j=1}^{m} e_{bj}$ is the expected number of deaths in group *b*, where

$$e_{bj} = \frac{n_{bj}\, d_j}{n_j}$$

and d_j, n_j are the total deaths and total number at risk in interval *j*, respectively. The test statistic is compared to a χ^2 distribution with *r* degrees of freedom.

2.3 Cox proportional hazards model

In Section 2.1, we discussed using the survival time and the status of censoring in the estimation of the survival function using the Kaplan Meier method. Non-parametric methods do not make specific assumptions about the distribution. We could assume some specific functional form for the hazard function and fit it to the data. We intend to consider incorporating the covariates of the individual into our analysis via the life time distribution. There are basically two ways to incorporate covariates into the analysis. These are the proportional hazard model (which is commonly used) and accelerated life models. A Cox proportional hazard model (Cox, 1972) is a statistical technique for exploring the relationship between the survival time and the covariates of a patient. It provides an estimate of the treatment effect on survival.

Suppose that for individual *i*, we have covariates

$X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,m})$, where $x_{i,k}$ is the k^{th} covariate value for the i^{th} individual for $k = 1,\ldots,m$ and individuals

$i = 1,.,n$. These covariates could be continuous or even indicator variables (equal to 1 if present and 0 if absent). We denote the hazard function of an individual *i* as $\alpha_i(t)$. In a proportional hazard model, we assume that for any two individuals *i* and *j*, the hazards are related by

$$\alpha_i(t) = \omega_{i,j} \,.\, \alpha_j(t) \; (2)$$

where $\omega_{i,j}$ is a constant that does not depend on *t*. We can rewrite Equation 2 as

$$\alpha_i(t) = \emptyset_i \,.\, \alpha_0(t) \; (3)$$

where $\emptyset_i$ is a constant which depends on the covariates of the individual *i* and $\alpha_0(t)$ is the baseline hazard function. The assumption is that the hazard $\alpha(t|X_i)$for an individual *i* with covariates X_i is

$$\alpha(t|X_i) = \alpha_0 e^{\beta X_i}$$

for some constants $\beta_{1,} \ldots\ldots, \beta_m$. The prognosis index or risk score is given as

$$\beta X_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_m x_{i,m}$$

3 Application

The data set originally included 8312 patients with 8172 females and 140 males (1.68%) patients with advanced breast cancer with ten covariates collected for 9-years (2004 to 2013) with the survival time, censoring indicator and cause of death. The covariates include age, marital status, gender, nationality, addresscode, topography, behaviour, grade, extent and laterality. All covariates were categorized except age which was continuous. There are 6 covariates which were completely observed for each patient: gender, age, laterality, nationality, topography and behaviour. The complete cases where all values for the covariates were recorded included 5432 patients. The rest of the unknown values will be regarded as missing values. The other covariates have at least one missing case. For the purpose of this research, analysis will be based on the complete cases. The mean age at diagnosis of the patients was 48.5 years with standard deviation of 12.57 and a range from 13 to 96. The number of males diagnosed of breast cancer in the complete data set was 68(1.25%) and females was 5364(98.75%). The ages of male in the data set have a range from 23 to 91 while female had a range from 13 to 96.

Kaplan Meier estimate allows us to estimate the survival function without assuming any particular model and hence it is non-parametric. The survival times amongst the different outcomes of a variable were compared using Kaplan Meier curves. Figure 1 shows the Kaplan.

Meier plot of data which can be used to estimate the survival probability at any time. The median time which gives the largest time for the survival estimate is 0.5 or higher is 4.96 years or approximately 5 years.

We can compare Kaplan Meier estimate of the survival distribution with that of the parametric fits.

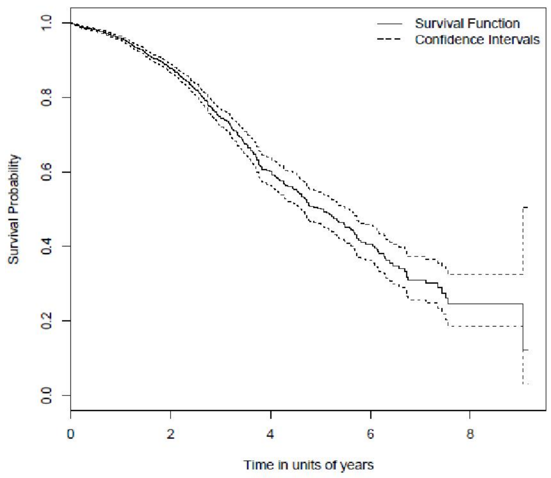


Figure 1: Kaplan-Meier survival curve for BC data, the dotted lines indicate the lower/upper bounds of the associated 95 confidence interval.

In our case, we use the exponential distribution where the survival function *S (t) = exp (-λt)*. We get a standard error of 0.0368 and the intercept parameter, *β* as 2.53. An approximate 95% confidence interval for *λ* is ($e^{-2.45}$, $e^{-2.60}$) = (0.074, 0.086). We can get estimates for 5 year survival probability, with limits as (0.65, 0.69).

We checked for significant differences between groups using log-rank test. We identified the factors that are significant in predicting a patients survival using the proportional hazard models. Figure 2 shows the Kaplan Meier estimates of the breast cancer survival by gender.

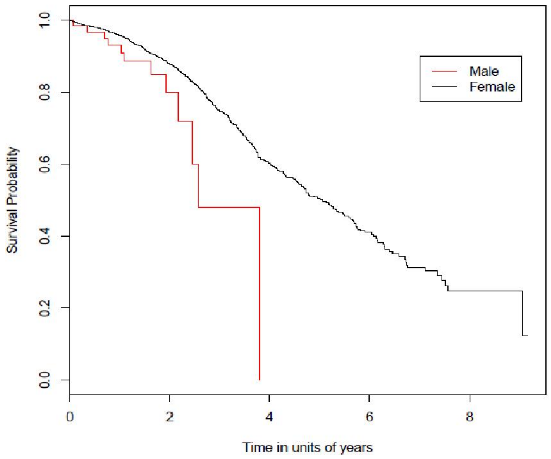


Figure 2: Kaplan-Meier estimates by gender

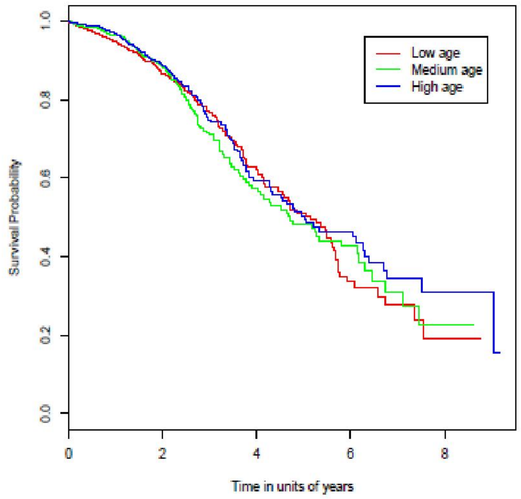


Figure 3: Kaplan-Meier estimates by age, categorised into high (≥50), medium (40 - 49), low (< 40)

We might want to investigate how time to event is influenced by covariates. In our models, we assume that covariates affect survival through a linear function βX_i (β is a vector of covariate effects and the prognostic index; X_i is the covariate vector of individual i). We assume a lognormal model and investigate the effects of the covariates on survival times in the BC data. The covariate effects are given in Table 1. The fitted survival model is given as

$$S(t) = 1 - \emptyset\left(\frac{logt - \beta x_i}{v}\right)$$

where *v* is the scale and it is 1.43. We get the survival probabilities of the individual depending on the covariates vector. We use the χ^2 test with null hypothesis that none of the covariates has any effect on survival. Since the *p-value = 0* which is small, we reject the null hypothesis and conclude that at least the covariates are important. We see that age, grade, marital status, address code, extent and topography are important and have effect on the survival but gender and laterality may not have. The positive coefficient for sex means that increasing the variable from 0 to 1 or changing sex from male to female leads to a decrease in hazard and so females do better than males, though not by statistically significant amount. We fit a Cox proportional hazard model using the covariates and the covariate effects are given in Table 2.

Table 2 shows that there are highly significant effects of at least some covariates. By inspection, laterality is not significant and important. Age is slightly significant. Increasing age leads to increased risk at a rate of 1% a year. Grade, address code, Extent and topography are highly significant. High grade, extent and topography counts as a risk factor. Moving a grade, extent or topography increases the risk by 49%, 62% and 7% respectively. Females do better with a hazard only about 0.44 the male value. So, females have a lower hazard than males. The address code is an important factor: patients in one address code compared to the next address code do better with hazard only about 0.95 of the former address code. Single, widowed or divorced people have worse prognosis than married people. Knowing the extent or stage of your breast cancer helps plan your treatment. The extent of breast cancer is the most important factor for prognosis. In general, the earlier the stage, the better the prognosis will be. In order to check whether a categorical covariate has a proportional effects on the hazards, we can re-fit a stratified model.

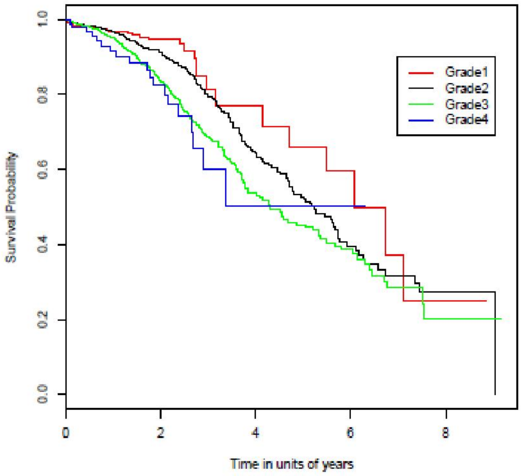


Figure 4: Kaplan-Meier estimates by Grade.

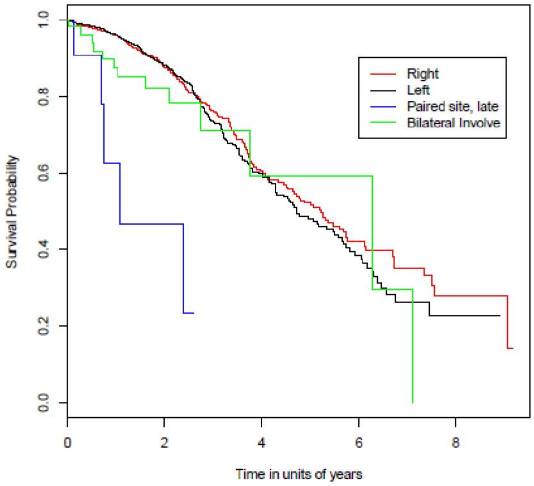


Figure 5: Kaplan-Meier estimates by Laterality.

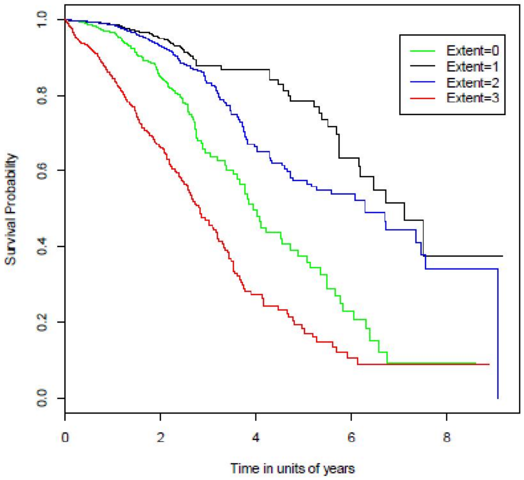


Figure 6: Kaplan-Meier estimates by Extent, Extent=0 indicate In situ, Extent=1 indicate Localised, Extent=2 indicate Regional NOS, Extent=2 indicate Regional: Dir Ext and Lymph node, Extent=2 indicate Regional: Direct Ext, Extent=2 indicate Regional: Lymph Node, Extent=3 indicate Distant Metastasis,

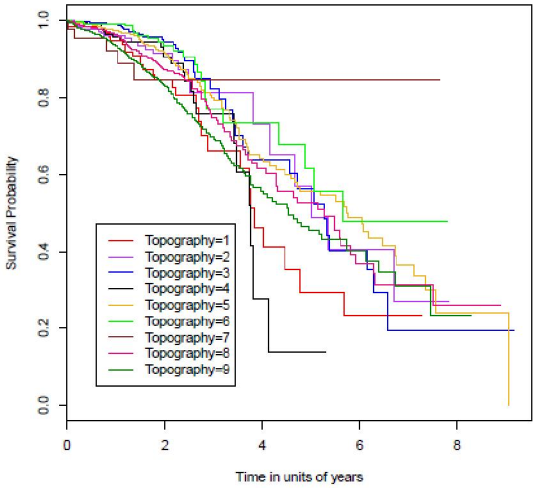
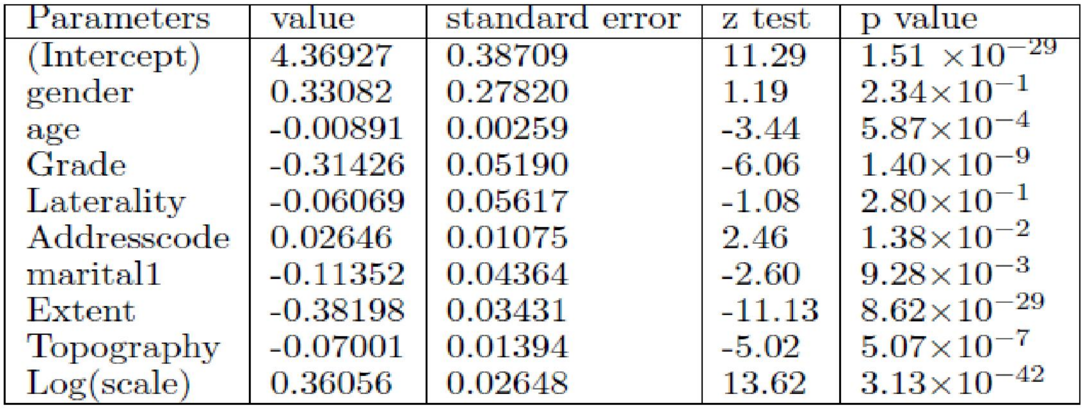


Figure 7: Kaplan-Meier estimates by Topography. Topography = 1, indicate C50.0 Nipple, Topography = 2, indicate C50.1 Central portion of breast, Topography = 3, indicate C50.2 Upper-inner quadrant of breast, Topography = 4 indicate C50.3 Lower-inner quadrant of breast, Topography = 5 C50.4 Upper-outer quadrant of breast, Topography = 6 indicate C50.5 Lower-outer quadrant of breast, Topography = 7 indicate C50.6 Axillary tail of breast, Topography = 8 indicate C50.8 Overl. lesion of breast, Topography = 9 indicate C50.9 Breast, NOS.

Table 1. Table showing the covariate effects using a lognormal model



Parameters	value	standard error	z test	p value
(Intercept)	4.36927	0.38709	11.29	1.51×10^{-29}
gender	0.33082	0.27820	1.19	2.34×10^{-1}
age	-0.00891	0.00259	-3.44	5.87×10^{-4}
Grade	-0.31426	0.05190	-6.06	1.40×10^{-9}
Laterality	-0.06069	0.05617	-1.08	2.80×10^{-1}
Addresscode	0.02646	0.01075	2.46	1.38×10^{-2}
marital1	-0.11352	0.04364	-2.60	9.28×10^{-3}
Extent	-0.38198	0.03431	-11.13	8.62×10^{-29}
Topography	-0.07001	0.01394	-5.02	5.07×10^{-7}
Log(scale)	0.36056	0.02648	13.62	3.13×10^{-42}

Table 2. Table showing the covariates using the Cox proportional hazard model

Parameters	coef	exp (coef)	se (coef)	Z	P (>|Z|)
gender	-0.8108	0.4441	0.29456	-2.753	0.006**
age	0.0071	1.0072	0.0030	2.364	0.018 *
Grade	0.4018	1.4945	0.0586	6.853	0.00***
Laterality	0.0799	1.0831	0.0626	1.277	0.2018
Addresscode	-0.0480	0.9531	0.0128	-3.768	0.0002***
marital	0.1457	1.1568	0.0492	2.963	0.0031**
extent	0.4852	1.6245	0.0437	11.127	0.000***
Topography	0.0712	1.0738	0.0163	4.362	0.00001***

4 Conclusion

Survival analysis is used to analyses data corresponding to survival time. Survival time is the time taken when an end event occurs in the data set. Additional, survival analysis provides special techniques that are required to compare the risks for death (or some other event) associated with different treatments or groups, where the risk changes over time. The most commonly used techniques are introduced in this work. Kaplan-Meier provides methods a statistical comparison of two groups, and Cox proportional hazard model. The work considered in this project was centered on survival analysis with application on Breast Cancer data.

Acknowledgements

We thank the Cancer Registry in King Faisal Specialist Hospital and Research Center (KFSH and RC) for providing the data for this study.

Corresponding Author:
Dr. Refah Alotaibi
Mathematical Science Department,
Princess Nourah Bint Abdulrahman University,
P.O. Box 84428, Riyadh 11671, Saudi Arabia.
E-mail: rmalotaibi@pnu.edu.sa

References

1. Al-Qahtani, M. S. (2007). Gut metastasis from breast carcinoma. Saudi Medical Journal,28:1590{1592. Smith MD, Wilcox JC, Kelly T, Knapp AK. Dominance not richness determines invasibility of tallgrass prairie. Oikos 2004;106(2):253–62.
2. American Cancer Society. (2013). Breast Cancer. http://www.cancer.org/cancer/
3. Chia, K. S., Du, W. B., Sankaranarayanan, R., Sankila, R., Wang, H., Lee, J., Seow, A., and Lee, H. P. (2004). Do younger female breast cancer patients have a poorer prognosis? Results from a population-based survival analysis. International journal of cancer, 108:761-765.
4. Chung, M., Chang, H. R., Bland, K. I., and Wanebo, H. J. (1996). Younger women with breast carcinoma have a poorer prognosis than older women. Cancer, 77:97-103.
5. Colleoni, M., Rotmensz, N., Robertson, C., Orlando, L., Viale, G., Renne, G., Luini, A., Veronesi, P., Intra, M., and Orecchia, R. (2002). Very young women (less than 35 years) with operable breast cancer: features of disease at presentation. Ann Oncol, 13:273-279.

6. Cox, D. R. (1972). Regression models and life tables (with discussion). J. R. Statist. Soc. B, 34:187-220.
7. Elkum, N., Dermime, S., Ajarim, D., Al-Zahrani, A., and Alsayed, A. (2007). Being 40 or younger is an independent risk factor for relapse in operable breast cancer patients: the Saudi Arabia experience. BMC Cancer, 7:222.
8. El Saghir, N. S., Seoud, M., Khalil, M. K., Charafeddine, M., Salem, Z. K., and Geara, F. B. (2006). Effects of young age at presentation on survival in breast cancer. BMC Cancer, 6:194.
9. Han, W., Kim, S. W., Park, I. A., Kang, D., Kim, S. W., Youn, Y. K., Oh, S. K., Choe, K. J., and Noh, D. Y. (2004). Young age: an independent risk factor for disease-free survival in women with operable breast cancer. BMC cancer, 4:82.
10. Ibrahim, E. M., Zeeneldin, A. A., Sadiq, B. B., and Ezzat, A. A. (2008). The present and the future of breast cancer burden in the kingdom of Saudi Arabia. Med Oncol, 25:387-393.
11. Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomp lete observations. J Am Statist Assn, 53:457-481.
12. Khan r, A., Frikha, M., Kallel, F., Meziou, M., Trabelsi, K., Boudawara, T., and Mnif, J. (2006). Breast cancer in young women in the south of tunisia. Cancer Radiother,10:565-571.
13. Lee, E. T. and Wang, J. W. (2003). Statistical Methods for Survival Data Analysis. Wiley.
14. Lozano, R., Naghavi, M. and Foreman, K., Lim, S., Shibuya, K., and Aboyans, K. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. Lancet, 380:2095-2128.
15. Mokdad, A. H., Jaber, S., Aziz, M. I., Al Buhairan, F., Al Ghaithi, A., and Al Hamad, N. M. (2014). The state of health in the arab world, 19902010: an analysis of the burden of diseases, injuries, and risk factors. Lancet, 383:309-320. 16. Nixon, A. J., Neuburg, D., Hayes, D. F., Gelman, R., Connolly, J. L., and Schnitt, S. (1994). Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage i and stage ii breast cancer. J Clin Oncol, 12:888-894.
16. Registry, N. (2007). Cancer Incidence Report Saudi Arabia 2002. Riyadh.
17. Robson, M., Gilewski, T., Haas, B., Levin, D., Borgen, P., Rajan, P., Hirschaut, Y., Press- man, P., Rosen, P. P., and Lesser, M. L. (1998). Brca-associated breast cancer in youngwomen. J Clin Oncol, 16:1642-1649.
18. Tarone, R. E. (2006). Breast cancer trends among young women in the united states. Epidemiology (Cambridge, Mass), 17:588-590.
19. Zabicki, k., Colbert, J. A., Dominguez, F. J., Gadd, M. A., and Hughes, K. S. and Jones, J. L. (2006). Breast cancer diagnosis in women 40 versus 50 to 60 years: increasing size and stage disparity compared with older women over time. Ann Surg Oncol, 13:1072-1077.

6/22/2017