Open Access Journal of Computer Science and Engineering
Volume 1, Issue 1, 2024, PP:14-21
www.aytinpublications.com

Research Article

Survival Analysis of Patients Who Had Undergone Surgery for Breast Cancer Using Cox Model

Bohang Li[1], Su Diao[2]*

[1]*Department of data science, Shopee Pte. Ltd., 118265, Singapore, Email: libohang.pku.edu.cn*
[2]*Department of Industrial & Systems Engineering, Auburn University, Alabama, USA, 36849*

***Corresponding Author:** Su Diao, Department of Industrial & Systems Engineering, Auburn University, Alabama, USA.*

Abstract

Breast cancer is the second leading cause of death in women, despite breakthroughs in cancer treatment. Breast cancer is the most commonly diagnosed malignant tumour in Western women, affecting roughly 1.7 million people worldwide each year. One of the most often used strategies for assessing survival data is the Cox model. However, parametric approaches may yield better predictions in some cases. A Weibull parametric model was used to analyse possible prognostic markers that may affect the survival of breast cancer patients in this study. The association between breast cancer patient variables and time to death, and the determination of power in the cancer patient's conditional inference tree, was evaluated using a semi parametric survival analysis approach. This study presents a promising scenario for the use of immunotherapy to treat patients with HER2 over expression. More research might be done to see how successful immunotherapy is in patients with various diseases, in order to improve their prognosis and quality of life.

Keywords: Survival analysis, Breast cancer, Quality of life (QOL), Cox Mode, breast cancer patients, semi parametric survival analysis.

Aricle Information

Recieved: 08 August 2024

Accepted: 21 August 2024

Published: 03 September 2024

Cite this article as:

Li Bohang, Su Diao. Survival Analysis of Patients Who Had Undergone Surgery for Breast Cancer Using Cox Model. Open Access Journal of Computer Science and Engineering, 2024;1(1); 14-21.

Introduction

Breast cancer is the most common cancer diagnosis in women around the world and the second most common cause of cancer-related death in women. Although the overall survival rate after breast cancer surgery is good in the general population, a variety of factors, such as demographic and clinical characteristics, care quality, and quality of life (QOL) prior to surgery, can affect survival dramatically. As a result, the capacity to produce accurate 10-year survival estimates following breast cancer surgery can assist healthcare organisations be more efficient in allocating and organising scarce healthcare resources to treat these patients throughout time.

Using survival rates After a certain period of time (usually 5 years) after diagnosis, the proportion of people who are still alive with the same type and stage of cancer can be estimated. Though they are unable to predict your life expectancy, they may be able to provide you with a better understanding of the likelihood that your treatment will be successful.

It's critical to keep things in perspective. While survival rates can be predicted and are frequently based on the outcomes of large groups of people who have previously been diagnosed with a specific cancer, they cannot be used to forecast what will happen in any given case. These figures can be baffling, and they may raise more doubts in your mind. To find out if these figures apply to you, speak with your doctor who is knowledgeable with your condition.

What is the Relative Survival Rate Over a 5-Year Period?

The term "relative survival" is used to compare women with breast cancer of the same type and stage to women in the general population. If the 5-year relative survival rate for a particular stage of breast cancer is 90%, women with that cancer are 90% more likely to survive at least 5 years after diagnosis than women without cancer. It means high.

The survival rates for several cancer kinds are based on data from the National Cancer Institute's SEER database, which is updated on a regular basis (NCI). The American

Cancer Society uses this data to calculate survival rates for various forms of cancer.

This database monitors breast cancer patients' 5-year relative survival rates in the United States, which are determined based on how far the illness has spread. The SEER database, on the other hand, does not use the AJCC TNM stages to categorise malignancies (stage 1, stage 2, stage 3, etc.). Cancers are divided into three stages: localised, regional, and distant.

The disease has remained localised: no sign of the cancer spreading outside the breast tissue has been found. Regional cancer, as opposed to locally progressed cancer, refers to cancer that has migrated outside of the breast to surrounding structures or lymph nodes.

Long-distance spread: The cancer has spread to other organs, including the lungs, liver, and bones.

Data Description

Age is the patient's age at the time of surgery, and Operation year represents the patient's surgery year (year - 19XX). The term Nb pos detected refers to the number of positive axillary nodes found. Survival status is abbreviated as Surv (class attribute) 1 indicates that the patient has lived for at least 5 years. 2 indicate that the patient died within 5 years.

Table 1.

	Age	Operation year	Nb_pos_detected	Surv
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1

Panda Profiling

Pandas profiling is an open source Python module that allows you to perform exploratory data analysis in minutes with just a few lines of code. Furthermore, if that isn't enough to persuade you to utilise this tool, it also creates interactive reports in web format that can be shown to anyone, even if they have no programming experience.

Dataset Info

Table 2.

Number of variables	4
Number of observations	306
Total Missing (%)	0.00%
Total size in memory	9.6 KiB
Average record size in memory	32.3 B

Variables Types

Table 3.

Numeric	3
Categorical	0
Boolean	1
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

Exploratory Data Analysis on Haberman's Cancer Survival Data Set

Haber Man's Survival Dataset

The dataset contains examples of studies on the survival of patients who underwent breast cancer surgery at the University of Chicago's Billings Hospital between 1958 and 1970.

The following are the data set's various attributes:

- Patient's age at the time of surgery (numerical)
- The year of the patient's procedure (year between 1958 to 1970, numerical)
- The number of axillary nodes that have been found to be positive (numerical)
- Survival status (a class attribute) is denoted by the following:
- 1 — if the patient survives for more than 5 years
- 2 — if the patient dies within 5 years

Objective

Based on the attributes, Perform exploratory data analysis to determine if a patient will live for more than 5 years.

EDA Model Analysis

EDA strategy is exactly that: strategy. It's more about the idea and philosophy of how data analysis should be done, rather than a collection of technologies. EDA is not the same as statistical graphics, but these terms are sometimes used interchangeably. Statistical graphs are a category of data characterization techniques that are all visual-based and focus on a single component of data characterization. EDA is a collective term for data analysis processes that avoids assumptions about what type of model the data follows, and instead allows the data to directly reveal its underlying structure and model. EDA is a way to think about how to analyze a collection of data, what to look for, how to search, and how to interpret it. EDA makes heavy use of a set of techniques called "statistical graphs," but statistical graphs are not the same.

Most of the EDA methods are visual and some quantitative techniques have been added for security. One of the reasons for the emphasis on graphs is that one of EDA's main missions is to openly investigate, providing analysts with unparalleled capabilities and convincing data to reveal structural secrets. Reveal and always be new, often unexpectedly prepared. Insight into the data. The combination of graphics and natural pattern recognition gives you unparalleled power in this regard.

Univar Ate Analysis

A fundamental statistical data analysis technique is univariate analysis. There is just one variable in the data, you don't have to work on causality. For example, consider a classroom survey. Researchers want to know how many men and women are in the room. The information provided here is limited to a single variable, number, and amount of variables. The main purpose of univariate analysis is to explain the data to find patterns. This is done using mean, mode, median, standard deviation, spread, and other statistics.

The most basic type of data analysis is univariate analysis. Uni stands for "one," meaning that there is just one sort of variable in the data. Univariate analysis' main purpose is to characterize the data. The information will be collected, analyzed, and a pattern will be discovered.

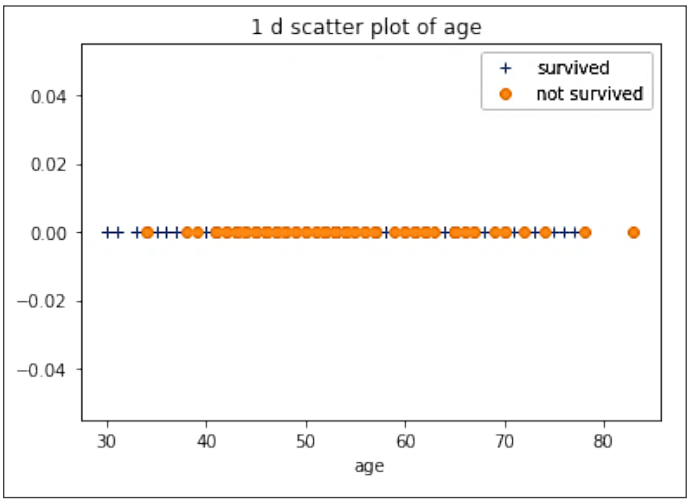

Figure 1. *OBSERVATION: Patients with age less than 33 survived. Patients with age greater than 80 could not survive.*

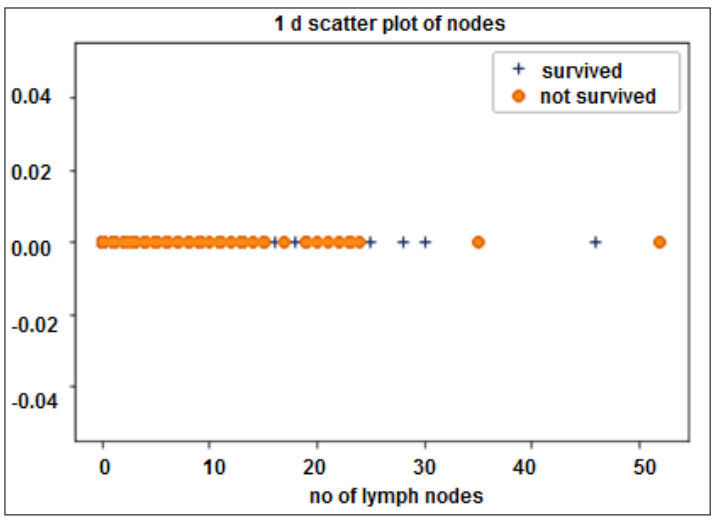

Figure 2. *Observation: a). Lymph nodes greater than 50 cannot survive*

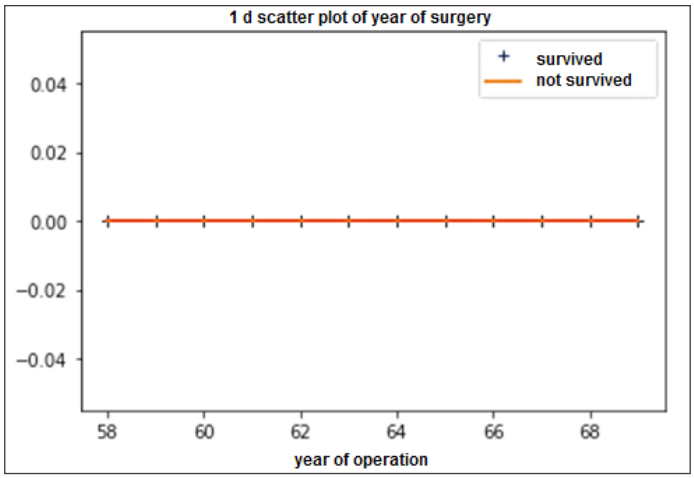

Figure 3. *The graph is observing highly overlapping data*

As we can observe from the above analysis that the EDA model of analysis is not able to derive any defined results in our analysis. Hence we are now moving to another model for the survival analysis.

Survival Analysis

The survival analysis is based on The probability that the event of interest did not occur at time t. A Survival function depending on time S(t) is generally used to represent that probability.

$$S(t) = P(T > t) S(t) = P(T > t) \quad (1)$$

To put it another way, S (t) is the probability of survival after time t. Lifespan randomly selected from the population is represented by T. S (t) is a decreasing function of t, and it ranges from zero to one (inclusive).

The Hazard Function

The hazard function is defined as the probability that a subject will experience an interesting event within a short time frame, assuming that the individual survives to the beginning of that interval. It is the instantaneous velocity multiplied by the time interval and is considered constant. Or you can think of it as a chance to encounter an interesting event at time t. It is the number of subjects alive at time t and the width of the interval divided by the number of subjects with events at intervals starting at time t.

This is because continuous random variables are unlikely to reach a certain value. That is why we assess the chance of an event occurring at a specific time interval between T and (T+T). Because our goal is to determine the risk of an occurrence, we do not want the risk to increase as the time interval T increases. As a result, we divide the equation by T to account for this. The equation is now scaled by T. The Hazard Rate equation is as follows:

$$h(t) = \lim_{\delta t \to 0} \frac{pr(t \leq Tt + \delta t \backslash T > t)}{\delta t} \quad (2)$$

The fact that T approaches 0 indicates that it is our goal to assess the probability of an event occurring at a particular point in time. As a result, moving the limit T closer to zero creates a very short period of time. It is important to note that hazards are not probabilities. This is because the denominator T can be a number greater than 1 while the probability is in the numerator.

Censored Data

When information concerning an observation's survival time is lacking, it is referred to as suppressed. This occurs when the subject is no longer noticed before the event of interest occurs. This can be due either to the end of the observation period or to the occurrence of another event preventing the observation of the event of interest. In an insurance context, lapsing leads to censored information when the event of interest is whatever mortality/morbidity risk. It is called right censored data. There are also left censored, interval censoring and left truncation but we will only work with right censored data. Someone who withdraws from a study before the end of the observation period and does not witness the occurrence is said to be right filtered.

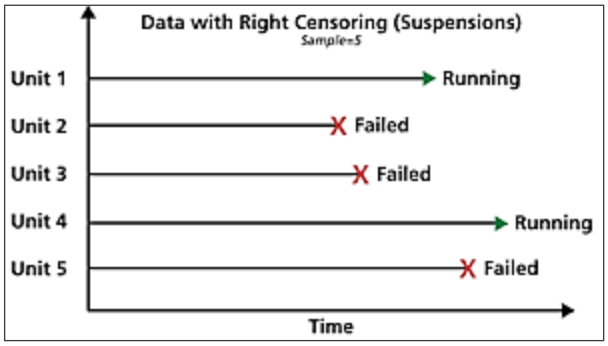

Figure 4. *This illustration, Units 1 and 4 did not experience the event before the investigation was completed.*

Data Analysis

The Kaplan Meier Estimator

The survival function is estimated using the Kaplan Meier estimator. Calculate the percentage of participants who survived a particular time (t).Before using KM estimate in survival analysis, three assumptions must be made:

- Subjects who are censored have the same chance of surviving as those who are followed.
- Regardless of when they are enrolled in the trial, all individuals have the same chance of survival.
- The target event will occur at the specified time. This is because the incident may have occurred between the two exams. If inspections are done frequently, i. H. The expected survival time can be measured more accurately if the time interval between tests is minimal.

The Kaplan-Meier curve is the most common visual representation of this function. It depicts the likelihood of an event occurring at a specific time interval (For instance, survival). The curve should approximate the genuine survival function for the population examined if the sample size is large enough. Dividing the number of surviving subjects by the number of people at risk is equal to the probability of survival for a particular period of time. The denominator does not include the censored subjects. The following is the equation:

$$S(t) = \prod_{t_i,t} \frac{n_i - d_i}{n} i \qquad (3)$$

The number of people at risk before time t is represented by ni, and the number of events of interest at time t is represented by di.

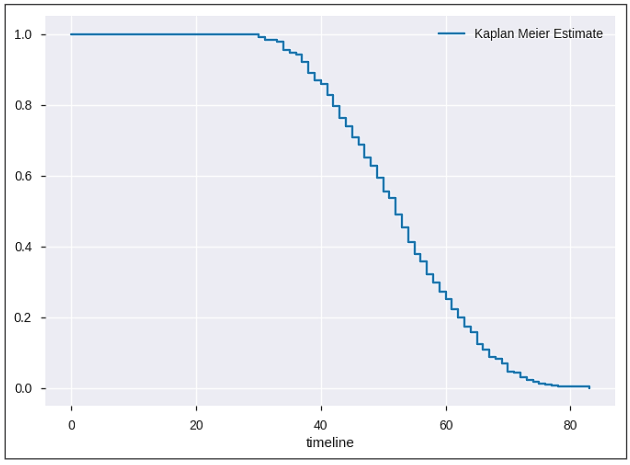

Figure 5.

The chance that the subject hasn't encountered Events of interest after time t, where time t is on the x-axis, are represented on the y-axis. The confidence intervals are used to determine how certain we are about the point estimations. The mid-term is when half of the population experiences an event of interest on average.

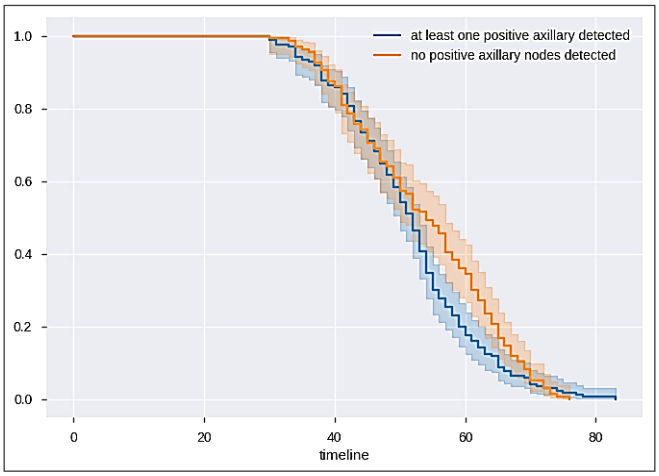

Figure 6.

There are two survival curves, one for each cohort. From the curves, it is evident that patients, who have undergone at least 1 surgery, are more likely to die sooner. Before 45 years old, the two curves are almost overlapped. At any point t>45 across the timeline, we can see that the survival probability of the cohort in blue is less than the cohort in red.

Cox Proportional Hazards model

Survival Regression

Survival regression uses additional functionality and length, and censored variables as covariates. These covariates are "regressed" with periodic variables. The survival regression dataset must be in the form of a Data

Frame that contains a column of time of interest, a column of whether or not the event of interest has occurred, and other covariates that need to be regressed. As with any regression algorithm, you need to prepare the data before feeding it to your model.

Cox Model

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Cox proposed the Cox Proportional Hazards Model, which considers the effects of numerous variables at once and investigates the relationship Between the survival distribution and these variables. This is similar to multiple regression analysis, but the dependent variable is a hazard function at a particular time t. It is based on a very short time interval that contain only one event of interest. In a Proportional Hazard Model, it is a semi-parametric approach for estimating weights. The weights' partial likelihood is maximised to get the parameter estimations. The steepest descent method is used to fit the Cox model to the data. Gradient Descent explains how to determine the weights that minimise the mistake. The Cox Proportional Hazards Regression Model has the following formula. The model is designed in order that an character subject`s log-hazard is a linear function in their static covariates and a converting population-stage baseline hazard function. Partial likelihood can be used to estimate these covariates.

$$h(t|x) = b_0(t)\, exp \sum_{i=1}^{n} b_i\ (x_i) \qquad (3)$$

- T represents survival time and risk can change over time.
- H (t) is a hazard function determined by a set of n covariates (x1, x2... xn).
- $b_0(t)$ When all other covariates are zero, The basic hazard function is defined as the probability of encountering the event of interest. If all xi are zero (size exp (0) = 1), then the hazard is 1. This is the only time-dependent component of the model. The model assumes a parametric form of the effect of covariates on hazazrds, not baseline hazard functions.
- exp∑ni=1bi (xi) is a potential threat is a scalar factor with no time dependence that solely will increase or decreases the baseline risk. In normal regression, it`s much like the intercept. The variables, or regression coefficients x, constitute the proportional alternate within side the risk that may be expected.
- The coefficients (b1, b2,..., bn) degree the impact (i.e., the impact size) of covariates.

The sign of the regression coefficients,, influences a subject's hazard. The baseline hazard will increase or decrease if these regression coefficients or variables change. A positive sign of $ b i $ indicates that the risk of an event is high and therefore the probability of the event of interest is higher for that subject. Negative sign, on the other hand, indicates that the event's risk is smaller. It's also worth noting that the size, or the value itself, has a factor. For example, a value of one for a variable indicates that it will have no effect on the Hazard. It will reduce the Hazard if the value is less than one, and it will increase the Hazard if the value is larger than one. The partial likelihood is used to estimate these regression coefficients, b.

The baseline hazard function does not need to be specified in the Cox proportional hazards model, so it can be modified and new parameters are available for each unique lifetime. However, the odds ratio is expected to remain proportional over the considered period. This improves the adaptability of the model. The basic hazard function can be parameterized according to a specific model of survival time distribution in the full parametric proportional hazards model. The Cox model can process right censored data directly, but it cannot process left censored or interval censored data.

The Cox Model is based on three assumptions:

- The Hazard Ratio of two topics is still the same.
- The Hazard Function's Explanatory Variables
- Individual participants' failure times are unrelated to one another.

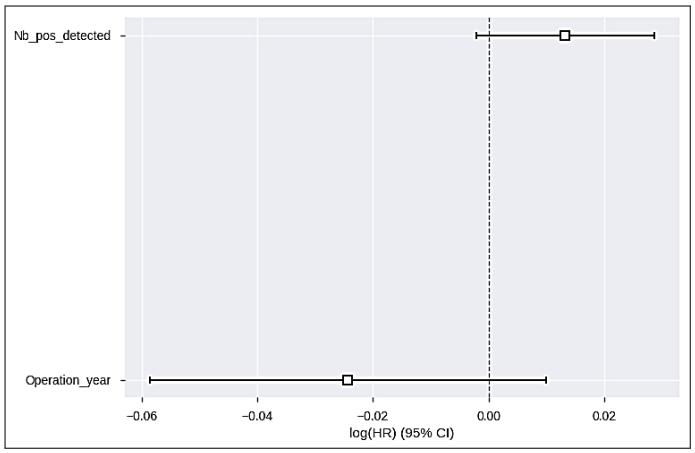

Figure 7.

The summary statistics above indicates the significance of the covariates in predicting the Survival risk both features play a tiny significant role in predicting the survival. The large CI indicates that more data are needed.

Table 4. *Example of a Survival Curve*

	Operation _year	Nb_pos_detected
4	65	4
125	64	0
211	67	0

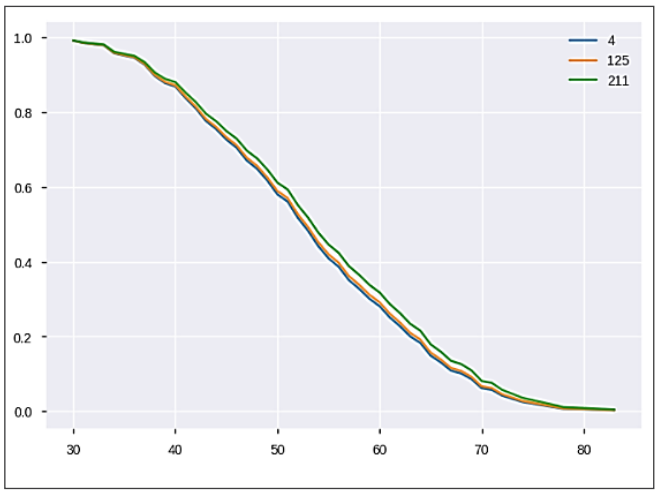

Figure 8.

The survival curves at the customer level are displayed. It depicts the survival curves for three patients who were chosen at random.

Discussion

Compared to Cox regression, decision tree analysis survival models have several advantages, including explicit maximisation of predicted accuracy, parsimony, statistical resilience, and transparency. As a result, researchers who want to build models with exact predictions and explicit choice criteria should use the classification tree survival architecture. Weathers used predictive error curves and coincidence indices to determine if the three methods fit into five publicly available datasets. One of the drawbacks of this study is the lack of a diagnosis date, which creates a day's gap between diagnosis and hospitalization. In addition, the histological features of the tumour were ignored. This may have accelerated the mortality rate of certain women by showing tumours that are more aggressive or resistant to chemotherapy. This status can help extend the findings.

References

1. Akaike H (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19, 716-23.
2. Akbari A, Razzaghi Z, Homaee F, et al (2011). Parity and reastfeeding are preventive measures against breast cancer in Iranian women. Breast Cancer, 18, 51-5.
3. Akbari ME, Khayamzadeh M, Khoshnevis S, et al (2012). Five and ten years survival in breast cancer patients mastectomies vs. breast conserving surgeries personal experience. Iranian Journal of Cancer Prevention, 1, 53-6.
4. Akram M, Ullah MA, Taj R (2007). Survival analysis of cancer patients using parametric and non-parametric approaches. Pakistan Veterinary Journal, 27, 194.
5. Alizadeh A, Mohammadpour RA, Barzegar MR, et al (2013). Comparing cox model and parametric models in estimating the survival rate of patients with prostate cancer on radiation therapy. Journal of Mazandaran University of Medical Sciences (JMUMS), 23.
6. Altman D, De Stavola B, Love S, et al (1995). Review of survival analyses published in cancer journals. British Journal of Cancer, 72, 511
7. American Cancer Society 2014. Cancer Facts & Figures 2014. American Cancer Society (ACS) Atlanta, GA: American Cancer Society, 2014. 72 p., pdf.
8. Arpino G, Bardou VJ, Clark GM, et al (2004). Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. Breast Cancer Res, 6, 149-56.
9. Bland JM, Altman DG (2004). The logrank test. Bmj, 328, 1073
10. Bray F, Ren JS, Masuyer E, et al (2013). Global estimates of cancer prevalence for 27 sites in the adult population in 2008. International Journal of Cancer, 132, 1133-45.
11. Carroll KJ (2003). On the use and utility of the Weibull

model in the analysis of survival data. Controlled clinical trials, 24, 682-701.

12. Cox DR, Oakes D 1984. Analysis of survival data, CRC Press.
13. Eccles D, Simmonds P, Goddard J, et al (2001). Familial breast cancer: an investigation into the outcome of treatment for early stage disease. Familial cancer, 1, 65-72.
14. Eivazi-Ziaei J, Sanaat Z, Asvadi I, et al (2013). Survival analysis of breast cancer patients in northwest Iran. Asian Pac J Cancer Prev, 14, 39-
15. Faradmal J, Talebi A, Rezaianzadeh A, et al (2012). Survival analysis of breast cancer patients using cox and frailty models. Journal of Research in Health Sciences, 12, 127-30.
16. Ferlay J, Soerjomataram I, Ervik M, et al (2014). GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer; 2013.
17. Gonzalez AnguloAM, Broglio K, Kau SW, et al (2005). Women age ≤≤ 35 years with primary breast carcinoma. Cancer, 103, 2466-72.
18. Harris EE, Schultz DJ, Peters CA, et al (2000). Relationship of family history and outcome after breast conservation therapy in women with ductal carcinoma in situ of the breast. International Journal of Radiation Oncology* Biology* Physics, 48, 933-41
19. Hashemian AH, Beiranvand B, Rezaei M, et al (2013). A comparison between cox regression and parametric methods in analyzing kidney transplant survival. World Applied Sciences Journal, 26, 502-7.
20. HAYAT EA, Suner A, Burak U, et al (2010). Comparison of five survival models: Breast cancer registry data from ege university cancer research center. TurkiyeKlinikleri Journal of Medical Sciences, 30, 1665-74.
21. IHME 2013. The Global Burden of Disease: Generating Evidence, Guiding Policy, University of Washington Seattle, WA, USA.
22. Klein JP, Moeschberger ML 2003. Survival analysis: techniques for censored and truncated data, Springer Science & Business Media.
23. Kleinbaum DG, Klein M 2011. Survival Analysis: A Self-Learning Text, Third Edition, Springer.
24. Kuru B, Camlibel M, Ali Gulcelik M, et al (2003). Prognostic factors affecting survival and disease free survival in lymph node negative breast carcinomas. Journal of surgical oncology, 83, 167-72.
25. May AM, Struijk EA, Fransen HP, et al (2015). The impact of a healthy lifestyle on Disability-Adjusted Life Years: a prospective cohort study. BMC medicine, 13, 39.
26. Sun, Y. (2024). TransTARec: Time-Adaptive Translating Embedding Model for Next POI Recommendation. arXiv preprint arXiv:2404.07096.