Performance Evaluation of Distributional Models to Analyze Random Right-Censored Breast Cancer Failure Time Data

Madiha Liaqat
University of the Punjab
Shahid Kamal
University of the Punjab
Florian Fischer (florian.fischer@rwu.de)
Ravensburg-Weingarten University of Applied Sciences https://orcid.org/0000-0002-4388-1245
Waqas Fazil
Insitute of Nuclear Medicine & Oncology Lahore

Research article

Keywords: censoring, time to failure analysis, parametric models, accelerated failure time, Akaike Information Criterion, oncology

Posted Date: June 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-35024/v1

Performance Evaluation of Distributional Models to Analyze Random Right-Censored Breast Cancer Failure Time Data

Madiha Liaqat[1], Shahid Kamal[2], Florian Fischer[3] and Waqas Fazil[4]

[1]College of Statistical and Actuarial Sciences (CSAS)
University of the Punjab, Lahore, Pakistan.

[2]College of Statistical and Actuarial Sciences (CSAS)
University of the Punjab, Lahore, Pakistan.

[3]Institute of Gerontological Health Services and Nursing Research
Ravensburg-Weingarten University of Applied Sciences, Weingarten, Germany.

[4]Institute of Nuclear Medicine & Oncology Lahore (INMOL)
Lahore, Pakistan.

Corresponding author:
Dr. Florian Fischer
Ravensburg-Weingarten University of Applied Sciences
Institute of Gerontological Health Services and Nursing Research
Doggenriedstraße, 88250 Weingarten, Germany.
Email: florian.fischer@rwu.de
Phone: +49 751 501 9441

Abstract

Background: Censoring frequently occurs in disease data analysis. Typically, non-parametric and semi-parametric methods are used to deal with different types of censored data. Distributional random right-censored failure time models on breast cancer data are employed to empirically find out a best-fitted model. A large number of studies are available on complete and disease-free survival time, but very few have focused on time to death from breast cancer recurrence.

Methods: In this retrospective study, we investigated the impact of factors related to breast cancer on cause-specific failure time. We included data from women who suffered from breast cancer as a primary disease and observed recurrence. Several factors related to breast cancer incidence and prognosis are studied. A multivariate accelerated failure time (AFT) model is used to evaluate the combined effect of study factors on death due to breast cancer.

Results: Univariate Weibull model showed that all factors included in the model have a strong association with breast cancer failure time. These factors are age at diagnosis, age at recurrence, molecular markers (estrogen, progesterone receptors, and Her2.neu), tumor grade, chemotherapy, and radiotherapy. The best model for right-censored breast cancer failure time data was a Weibull AFT, which was chosen by a stepwise backward selection.

Conclusions: The AFT model is the best choice for the analysis of time to failure data when hazards are non-proportional, as it provides efficient estimates and an estimate of the median failure time ratios.

Keywords: censoring, time to failure analysis, parametric models, accelerated failure time, Akaike Information Criterion, oncology.

Background

Cancer causes a large disease burden. In 2017, worldwide 9.5 million deaths were recorded due to cancer overall, among which were about 600,000 cancer deaths due to breast cancer in women. Breast cancer is the most frequent cancer among women, leading to 1.9 million cases in 2017 with a significant increase in absolute cases during the past decades. Pakistan, being a lower-middle-income country, has a greater burden of breast cancer patients compared to its neighboring countries. It is even the country with the highest age-standardized death rate in 2017 globally [1]. Men and women both are vulnerable to breast cancer. However, men breast cancer rate is low, which is 1 in 1,000 over their lifetime, while women ratio is 1 in 8 [2]. An early recurrence of breast cancer in the first three to five years after primary treatment increases the probability of death [3].

Frequently, disease data are faced by incomplete information, called censoring. Three types of common censoring can be observed in study design: in type 1 and 11 all subjects under study have the same starting point, while in type 111 censoring subjects have different entering times in the predefined study period [4]. Censored data leads to biased estimates and a reduction in precision if analyzed with standard statistical techniques. In binary variables, 0 is used for censored data, while 1 indicates the occurrence of an understudy event [5].

Analysis of time to failure censored disease data is being used since the 1950s, for estimating the time span till a specific event [6]. Event time may be continuous or discrete, based on the timeframe window. Usually, non-parametric and semi-parametric methods are used as compared to parametric methods which have more assumptions to be fulfilled. Parametric methods produce reliable results in studies where the proportional hazard assumption is at stake [7]. Many

distributions can be used to define continuous non-negative failure time random variables. Among them, most frequently, are Weibull, exponential, log-normal, log-logistic, and Rayleigh [8].

By building a linear relationship between logarithm of failure time and covariates, data can be analyzed by an accelerated failure time (AFT) model, as illustrated by Lee and Go [9]. The increased incidence of breast cancer reported in Pakistani women led to the conduction of a large data analysis in order to investigate the potentially causal effects on breast cancer deaths, using both univariate and multivariate analysis. In this study, we used retrospective breast cancer data and applied AFT models by assuming different distributions towards the error term to discover the best model. The analysis has been conducted to find out the combined effect of factors that accelerated death of women after first breast cancer recurrence.

Methods

Study design

This study included a retrospective analysis of data from 1,028 primary breast cancer women, who have been observed recurrence after initial treatment from February 2011 to February 2018. Data were collected at the Institute of Nuclear Medicine & Oncology Lahore, Pakistan. Patients who were still alive (survived) at the end of the study or died from another cause than breast cancer, are considered right-censored.

Variables included in the study

While age has an influence on women's survival, Cianfrocca and Goldstein [10] studied the effect of prognostic factors in early stage breast cancer patients. Age at diagnosis and recurrence both were considered to elaborate the effect of age on death due to breast cancer in our study.

Estrogen receptors (ER), Progesterone receptors (PR), and Human epidermal growth factor receptor 2 (Her2.neu) were added as molecular markers in this study. Cancer is referred ER-positive if it has receptors for the hormone estrogen [11]. The same is for hormone progesterone (PR) [11,12]. Her2.neu directs the production of special proteins, which expresses higher aggressiveness and fast growth of cancer cells [12,13].

The histological grade is the reported degree of malignancy of the tumor. Grade 1 indicates a slight degree, while grades 11 and 111 show a severe malignancy [3]. With chemotherapy and radiotherapy, two initial treatment(s) assigned to patients were included. The indicator variable 1 indicated that the women have received the treatment.

Study failure time started after recurrence and lasted to death or end of follow up. Censorship event of patients "still alive" or "death other than breast cancer" are represented by the value 0, and 1 for patients who died because of breast cancer. The possible association between the understudy factors mentioned above and the time to death due to breast cancer after recurrence has been assessed using Weibull accelerated failure time model. This model was chosen because of its model fit after employing and comparing different parametric models [9].

Parametric models

Parametric models provide precise information the nature of data by fulfilling assumptions of randomness, normality, and homogeneity. We have performed parametric analysis to determine the best-fitted distribution that can outperform others on Pakistani women's breast cancer data. Our goal was to suggest an accurate model which can be used to predict breast cancer-specific death based on related factors. In this setting, the dichotomous response variable is time to failure denoted by symbol T ; death due to breast cancer and still alive or death other than breast cancer.

Here $Z = (Z_1, Z_2, Z_3, ...Z_p)$ are different predictors, related to breast cancer. The general relationship form between T and Z is given as

$$T = f(Z) + \varepsilon \quad (1)$$

Z is input variable, which may have impact on time to failure, and f is the unknown functional form. Practically, a 100% exact true relationship is not possible. To cover up uncontrolled chances of error, ε is also included in model. The linear relationship between f and Z has the following form:

$$f(Z) = \mu + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3, ...\beta_p Z_p \quad (2)$$

Patients' failure time T is random and non-negative, followed any specific distribution, with probability density function (pdf),

$$f(t) = -\frac{d}{d(t)} S(t) \quad (3)$$

where $S(t) = P(T > t)$ represents the survival function, which is the probability of a woman survivor up to a predefined time point $t\,(0 < t < \infty)$. While the hazard function

$$h(t) = \frac{\lim_{\Delta t \to 0} P\{T \in (t + \Delta t)\} / \Delta t}{S(t)} \quad (4)$$

is the probability of an instantaneous failure per unit of time given that an individual patient has survival after time t. $T = \min(Y, C)$ is failure time, while $\delta = I_{(Y \le C)} = (0 \ \ or \ \ 1)$ is an indicator for the event variable.

T failure time is followed any distribution, having probability density function (pdf)

$$f(t) = h(t)S(t) \quad (5)$$

After proposing the model, the maximum likelihood method is employed to get estimates of parameters [14]. Right-censoring general form of likelihood function is the product of density $f(t)$ and survival $S(t)$ function:

$$L = \prod_{j=1}^{p} f(t_j)^{\delta_j} S(t_j)^{1-\delta_j} \quad (6)$$

Here, p is the total number of patients which were observed for a specified period, and t_j is the number of patients who died due to breast cancer, for censoring indicator $\delta_j = 1$, The likelihood estimates are maximized using Newton Raphson procedure, which may be time consuming and tricky without computer programming. The freely available R software is used to implement these techniques [15].

Parametric models assume the shape of unknown functional form f, to estimate parameters $\mu, \beta_1, \beta_2, \beta_3, ..., \beta_P$. To get good estimates assuming parametric form should get close to true unknown f. The purpose of applying different flexible parametric models is to obtain different functional forms for f.

In lifetime distributions, Weibull is the most widely used with $\theta > 0$ shape and $\lambda > 0$ scale parameters [16,17]. Different parameterization forms of Weibull distribution have been used in literature, the simplest one for survival and hazard function is as follows:

$$f(t) = \theta\lambda t^{\lambda-1} \exp(-\theta t^{\lambda})$$

$$S(t) = P(T > t) = \exp(-\theta t^{\lambda})$$

$$h(t) = \frac{f(t)}{S(t)} = \theta\lambda t^{\lambda-1}$$

The exponential distribution is often concerned for experiments to account for the amount of time until a specific event occurs. The simplest distribution has one parameter $\theta > 0$ [18].

It is defined by distribution $f(t) = \theta \exp(-\theta t)$, survival $S(t) = P(T > t) = \exp(-\theta t)$, and hazard $h(t) = \frac{f(t)}{S(t)} = \theta$ function.

Log-normal is a probability distribution, with a normally distributed logarithm. It is widely used in lifetime data analysis [19,20]. Therefore, with scale $0 < \alpha < \infty$ and shape parameter $\beta > 0$, the general form of said distribution is

$$f(t) = \frac{\phi(\frac{\log(t) - \alpha}{\beta})}{t\beta}$$

$$S(t) = P(T > t) = \phi(\frac{\log(t) - \alpha}{\beta})$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{\phi(\frac{\log(t) - \alpha}{\beta})}{t\beta}}{\phi(\frac{\log(t) - \alpha}{\beta})}$$

Where $f(t)$ is the probability distribution function, $S(t), h(t)$ are survival and hazard functions, respectively.

Failure time $T \sim \text{loglogistic}(\theta, \lambda)$ follows the log-logistic distribution with a positive scale parameter θ and positive shape parameter λ [21]. The probability density function $f(t)$, survival function $S(t)$, and hazard function $h(t)$ are depicted in general form:

$$f(t) = \frac{\theta\lambda t^{\lambda-1}}{(1+\theta t^{\lambda})^2}$$

$$S(t) = P(T > t) = \frac{1}{1+\theta t^{\lambda}}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta\lambda t^{\lambda-1}}{1+\theta t^{\lambda}}$$

Rayleigh distribution having probability density, survival and hazard functions, $f(t) = \frac{2te^{-\frac{t^2}{\theta}}}{\theta}$ $S(t) = P(T > t) = e^{-\frac{t^2}{\theta}}$, $h(t) = \frac{f(t)}{S(t)} = \frac{2t}{\theta}$ respectively, is used to analyze lifetime data [22].

Accelerated failure time model

For evaluating the combined effect of covariates on time to death T_j, an AFT model is applied in the form of linear modeling:

$$\log T_j = \mu + \sum_{j=1}^{p} Z_j'\beta_j + W\varepsilon_j \qquad (7)$$

Where β_j are coefficients of unknown parameters, W is the scale parameter, and ε is the random error term [23]. All AFT models like Weibull, exponential, log-normal, log-logistic and Rayleigh are named for the distribution of T not for ε_j or $\log T$. The log form of the generalized model (7) can be written as,

$$\log(T_j) = \mu + \beta_1 Z_{j1} + \beta_2 Z_{j2} + \beta_3 Z_{j3} + \beta_4 Z_{j4} + \beta_5 Z_{j5} + \beta_6 Z_{j6} + \beta_7 Z_{j7} + \beta_8 Z_{j8} + W\varepsilon_j \qquad (8)$$

$$T_j = e^{\mu+\beta_1 Z_{j1}+\beta_2 Z_{j2}+\beta_3 Z_{j3}+\beta_4 Z_{j4}+\beta_5 Z_{j5}+\beta_6 Z_{j6}+\beta_7 Z_{j7}+\beta_8 Z_{j8}+W\varepsilon_j} \qquad (9)$$

Our parametric AFT models have the form

$$\log(Time) = \mu + \beta_1 Ageatdiagnosis + \beta_2 Ageatrecurrence + \beta_3 ER + \beta_4 PR + \beta_5 Her2.neu + \beta_6 Grade + \beta_7 Chemotherapy + \beta_8 Radiotherapy + W\varepsilon \qquad (10)$$

Where the unknown coefficients $\mu, \beta_1, \beta_2, \beta_3, ..., \beta_p$ and W in (10) are estimated by the method of maximum likelihood for patients who did not observe breast cancer cause-specific mortality,

$$L(\beta, \mu, W) = \prod_{j=1}^{p} \left\{ f_j(t_j) \right\}^{\delta_j} \left\{ S_j(t_j) \right\}^{1-\delta_j} \quad (11)$$

The log-likelihood function is given as

$$\log L(\beta, \mu, W) = \sum_{j=1}^{p} \left\{ -\delta_j \log(Wt_j + \delta_j \log f_{\varepsilon_j}(Z_j) + (1-\delta_j) \log \delta_{\varepsilon_j}(Z_j)) \right\} \quad (12)$$

where

$$Z_j = \frac{(\log T_j - \mu - \beta_1 Z_{ij} - \beta_2 Z_{2j} - \beta_3 Z_{3j} - ... - \beta_p Z_{pj})}{W} \quad (13)$$

Unknown parameters $\mu, W, \beta_1, \beta_2, \beta_3, ..., \beta_p$ are found by maximizing (12) using Newton Raphson procedure.

To compare parametric non-nested models, the Akaike information criterion (AIC) is used to select the best-fit model [24].

$$AIC = -2 \times \log(likelihood) + 2(P + Q) \quad (14)$$

Here, P represents the number of parameters and Q the number of the constant factor. Estimates of unknown parameters are obtained by employing the log-likelihood for censored data.

Results

In the present study, women's age was divided into two factors, which are age at diagnosis and at recurrence: the median age at diagnosis of breast cancer was 47 years (range: 18-59); while the median age at recurrence was 49 years (range: 21-62). Median survival time after recurrence was 3 years and just half (54.1%) of cancer was estrogen receptor negative. The majority of patients

were progesterone receptor positive (64.6%) and Her2.neu positive (52.9%). Overall, 207 women (20.1%) had tumor grade 1, whereas 821 (79.9%) had a higher level of malignancy. Chemotherapy (36.4%) and radiotherapy (87.4%) were given as primary treatments (Table 1).

There were 447 deaths among the 1,028 women. As shown in Table 1, 78.5% of deaths occurred due to breast cancer before three years after recurrence, 20.8% within 3 to less than 6 years, and 0.7% of patients died due to breast cancer after 6 years of its recurrence. The molecular markers among women who died due to breast cancer were distributed as follows: 59.3% ER-positive, 66.7% PR-negative, and 65.5% Her2.neu-positive. Breast cancer death was correlated positively with higher tumor grade (11 and 111; 97.3%) and no chemotherapy (67.8%).

For estimating the survival rate, the parametric Weibull and non-parametric (Kaplan-Meier) methods were used to draw survival probabilities against years of survival for breast cancer patients after recurrence (Figure 1). All of the variables listed at the outset were tested with the help of the Weibull accelerated failure model for death from breast cancer as the endpoint. Variables were first analyzed individually, that is, as main effects. Age at diagnosis $(P = 0.005)$ and age at recurrence $(P = 0.012)$ both have impact on disease-specific mortality. From the disease factors, estrogen receptor $(P = 1.0e^{-06})$, progesterone receptor $(P = <2e^{-16})$, and Her2.neu $(P = <2e^{-16})$ have proven their importance on women's life span after breast cancer recurrence. The study of tumor grade $(P = <2e^{-16})$ was important from the cause specific survival point of view. Both treatments chemotherapy $(P = 0.001)$ and radiotherapy $(P = 1.0e^{-06})$ showed the importance to predict deaths.

Table 2 describes the results of the multivariate accelerated failure time models we investigated if the covariates considered in the study are predictive of the primary breast cancer failure time for

206 women. To choose a suitable distributional model, we performed a preliminary analysis of the data
207 using Weibull, exponential, log-normal, log-logistic and Rayleigh AFT, by adding all factors in
208 the models. Based on these results, the Weibull AFT model had the best fit for the data under
209 study. We fitted models by the maximum likelihood method and the best-fit distributional model
210 was chosen based on the AIC criteria. AIC scores for all distributional models are shown in Table
211 3.
212 The final Weibull AFT model is simple and consisted only of relevant factors. For choosing the
213 best model, we applied backward stepwise factors selection (Table 4). Step by step we removed
214 all non-relevant factors and found the best model having factors as follows:

215 $$\log(Time) = \mu + \beta_4 PR + \beta_5 Her2.neu + \beta_6 Grade + \beta_7 Chemotherapy + \beta_8 Radiotherapy + W\varepsilon \quad (15)$$

216 $$S(\frac{T}{Z_j}) = e^{[Te^{-}(8.942+1.106PR-0.316Her2.neu-1.634Grade+0.190Chemotherapy-0.390Radiotherapy)]^{1.527}} \quad (16)$$

217 $$Z_j = \{Z_1 = PR, Z_2 = Her2.neu, Z_3 = Grade, Z_4 = Chemotherapy, Z_5 = Radiotherapy\}$$

218 All factors included in the final model impacted on event occurrence.
219 By considering the final model, the PR positive tumor reduces the risk of death as compared to the
220 PR negative tumor by 82% (HR=0.18, 95% CI: 0.15-0.22), Her2.neu positive patients showed a
221 greater hazard (HR=1.62, 95% CI: 1.33-1.97) of death due to breast cancer. An advanced tumor
222 grade (11 and 111) accelerated the risk of mortality (HR=12.19, 95% CI: 6.85-21.69. Women who
223 have gone through chemotherapy (HR=0.75, 95% CI: 0.61-0.91) experienced a longer life span
224 even after recurrence (HR=0.75, 95% CI: 0.61-0.91). However, radiotherapy status suggested no
225 reduction in time to death (HR=1.81, 95% CI: 1.26-2.62).
226
227

Discussion

In this paper, we applied specific parametric models to illustrate the important factors associated with breast cancer death and described the best-fitting model. The increased incidence of breast cancer reported in Pakistani women led to the conduction of a large data analysis in order to investigate the potentially causal effects on breast cancer deaths, using both univariate and multivariate analysis. We applied suitable distributional models to evaluate their performance based on breast cancer data.

Non-parametric and semi-parametric methods have been extensively used in previous research on disease data analysis, but computational complexity automatically kept parametric methods away. However, after advancement in computer programming, some researchers were able to make comparisons between semi-parametric and parametric methods [25]. Modeling time to death after recurrence in breast cancer survival is important from a clinical perspective, although, up to now, many studies have focused on mortality of breast cancer patients but very few considered time to failure after recurrence [3,11].

In the present study we estimated the factors which accelerate breast cancer mortality. AFT models make practical sense to study the influence of covariates on time to failure at a given significance level, although without knowing which distribution will be best-fitted for a case study. One simple solution is to check the assumptions of distribution for the data under study. Researchers designated specific case studies of diseases with different objectives to choose the best parametric model. The AFT model is the best choice for the analysis of time to failure data when hazards are non-proportional, as it provides efficient estimates and an estimate of the median failure time

ratios. From the interpretation point of view, AFT model's results are easy and help clinicians to make wise decisions related to the patients' conditions [23].

Distributional models are employed on data by including all factors to get the best model, based on different selection criteria [4,6,7,9]. According to our study, Weibull AFT is the best model to fit for right-censored breast cancer data after recurrence. We also used the best-fitted model, likelihood-based estimates to evaluate the combined association of different variables with breast cancer failure event (which is death). The results from the summarized data agreed with the already available evidence. Our findings suggested that PR negative, Her2.neu positive, higher tumor grade, and no chemotherapy increase the risk of death. The most surprising result is about radiotherapy treatment which depicted no reduction in breast cancer time to death. This might be due to a higher level of physical impairment of patients receiving radiotherapy treatment. However, patients treated by radiotherapy at an early stage have a larger survival time [26].

We did not include the interaction of chemotherapy and radiotherapy. Furthermore, other treatments should be included to find out treatments' effect. Information on treatment after recurrence has not been included, as many studies reported the effectiveness of chemotherapy after recurrence [3,27]. The relationship between breast cancer death and age has been a controversial topic [28,29]. Our investigation suggested no effect of age at diagnosis and recurrence on time to breast cancer death. This research provided an example of a situation where other time to failure techniques are inappropriate and the AFT model provides a better description of failure time. Non-parametric AFT models can also be used, because they require no specifications of the distribution. We recommend, to conduct further comparisons between parametric AFT and non-parametric AFT models.

Limitations

There are several limitations to our study. Firstly, this is a retrospective rather than a prospective study. Therefore, the quality and availability of information may limit the external validity of this study. Secondly, we investigated only limited available time independent factors. It would be of interest to investigate the effect of time dependent factors as well information about sites of recurrence was not added in this study, which may be helpful to provide better prediction about time to failure. Stage of cancer, tumor size, and involvement of lymph nodes could be added as input variables to receive valuable information about the patients' overall health status. Moreover, a mixture of two distributions can be used for better fit of data. Thirdly, to get in depth analysis of breast cancer data in the context of Pakistani women, overall survival should also be estimated. Finally, a simulation study can also be designed to check model assumptions.

Conclusions

The study on using random right-censored breast cancer data shows that the AFT model is the best choice for the analysis of time to failure data when hazards are non-proportional, as it provides efficient estimates and an estimate of the median failure time ratios.

List of abbreviations

AFT Accelerated failure time

AIC Akaike information criterion

CI Confidence interval

ER Estrogen receptor

HR Hazard ratio

PR Progesterone receptor

DECLARATIONS

Ethics approval and consent to participate

According to the Ethical Guidelines for Epidemiologists (IEF-EGE) and the regulations of the ethics committee located at the Advanced Studies and Review Board, University of the Punjab Lahore (Pakistan), no ethics approval is needed, because the analysis is based on routine data. At data collection, all patients provided written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

Data is available upon reasonable request from the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

The work was supported by the Higher Education Commission Pakistan under grant No. 46-2SS2-123 awarded to the first author.

Authors' contributions

The study was conceptualized by ML, supported by SK. WF has been responsible for data acquisition. ML analyzed the data, SK, FF and WF supervised this process. ML drafted the

manuscript, SK, FF and WF revised it critically for important intellectual content. All authors reviewed the final version of the manuscript.

Acknowledgements

We thank the staff of the Institute of Nuclear Medicine & Oncology Lahore (INMOL), who supported in data collection. We also wish to thank Dr. Rab Nawaz Maken from INMOL cancer hospital, Lahore, Pakistan, to provide full support to conduct this research.

References

1. IHME. Global Burden of Disease study 2017 – Visualization tool, GBD Compare. 2018. https://vizhub.healthdata.org/gbd-compare/. Accessed 5 June 2020
2. Anderson WF, Jatoi I, Tse J, Rosenberg PS. Male breast cancer: a population-based comparison with female breast cancer. J Clin Oncol. 2010;28:232–9.
3. Lafourcade A, His M, Baglietto L, Boutron RMC, Dossus L, Rondeau V. Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. BMC Cancer. 2018;18:171.
4. Lee E. Statistical Method for Survival Data Analysis. New York: Wiley; 1992.
5. Kleinbaum DG, Klein M. Survival Analysis – A Self-Learning Text. New York: Springer; 2005.
6. Hosmer DW, Lameshow S, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data. Hoboken: Wiley; 2008.
7. Cox DR, Oakes D. Analysis of Survival Data. London: Chapman and Hall/CRC; 1984.
8. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. Hoboken: Wiley; 2002.
9. Lee ET, Go OT. Survival analysis in public health research. Ann Rev Public Health. 1997;18:105–34.
10. Cianfrocca M, Goldstein LJ. Prognostic and predictive factors in early-stage breast cancer. Oncologist; 2004;9:606–16.
11. Knight WA, Livingston RB, Gregory EJ, McGuire WL. Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer. Cancer Res. 1977;37:4669–71.

12. Osborne CK, Yochmowitz MG, Knight WA. The value of estrogen and progesterone receptors in the treatment of breast cancer. Cancer. 1980;46:2884–8.
13. Perou CM. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.
14. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J R Stat Soc; 1949;11:15–53.
15. Pourhoseingholi MA, Hajizadeh E, Moghimi DB, Safaee A, Abadi A, Zali MR. Comparing Cox regression and parametric models for survival of patients with gastric carcinoma. Asian Pac J Cancer Prev. 2007;8:412–6.
16. Weibull W. A statistical distribution function of wide applicability. Journal of Applied Mechanics. 1951;18:293–7.
17. Peto R, Lee PN. Weibull distributions for continuous carcinogenesis experiments. Biometrics. 1973;29:457–70.
18. Benjamin E. Exponential Distribution and Its Role in Life Testing. Stanford: Wayne State & Stanford Universities; 1958.
19. Crow EL, Shimizu K. Log-Normal Distributions: Theory and Application. New York: Dekker; 1988.
20. Royston P. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. Statistica Neerlandica. 2011;55:89–104.
21. Voorn WJ. Characterization of the logistic and log-logistic distributions by extreme value related stability with random sample size. Journal of Applied Probability. 1987;24:838–51.
22. Lee WC, Wu JW, Hong CW, Hong SF. Assessing the lifetime performance index of Rayleigh products based on the bayesian estimation under progressive type II right censored samples. J Comput Appl Math. 2011;235:1676–88.
23. Wei LJ. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Stat Med. 1992;11:1871–9.
24. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19:716–23.
25. Klein JP, Zhang MJ. Survival Analysis Software. Hoboken: John Wiley & Sons; 2005.
26. Bhoo-Pathy N, Verkooijen HM, Wong FY, Pignol JP, Kwong A, Tan EY, et al. Prognostic role of adjuvant radiotherapy in triple-negative breast cancer: a historical cohort study. Int J Cancer. 2015;137:2504–12.

27. O'Rorke MA, Murray LJ, Brand JS, Bhoo-Pathy N. The value of adjuvant radiotherapy on survival and recurrence in triple-negative breast cancer: a systematic review and meta-analysis of 5507 patients. Cancer Treat Rev. 2016;47:12–21.
28. Hartley MC, McKinley BP, Rogers EA, Kalbaugh CA, Messich HS, Blackhurst DW, et al. Differential expression of prognostic factors and effect on survival in young (≤40) breast cancer patients: a case-control study. Am J Surg; 2006;72:1189–94.
29. Brenner H, Hakulinen T. Are patients diagnosed with breast cancer before age 50 years ever cured? J Clin Oncol; 2004;22:432–8.

Figure 1: Weibull survival plot

394 **Table 1:** Characteristics of factors associated with breast cancer diagnosis

Factors	***Death due to breast cancer (n=447)***	***Alive or death due to another disease than breast cancer (n=581)***	***Total (n=1028)***
Age at diagnosis (in years)			
Mean (SD)	44.0 (7.81)	45.6 (7.74)	44.9 (7.81)
Median [Min, Max]	47.0 [18.0, 59.0]	48.0 [22.0, 59.0]	47.0 [18.0, 59.0]
Age at recurrence (in years)			
Mean (SD)	46.2 (7.67)	47.3 (7.66)	46.9 (7.68)
Median [Min, Max]	49.0 [21.0, 61.0]	49.0 [24.0, 62.0]	49.0 [21.0, 62.0]
Survival time after recurrence (Years)			
0 to <3	351 (78.5%)	156 (26.9%)	507 (49.3%)
3 to <6	93 (20.8%)	378 (65.1%)	471 (45.8%)
≥6	3 (0.7%)	47 (8.1%)	50 (4.9%)
Estrogen receptor (ER)			
Negative	182 (40.7%)	374 (64.4%)	556 (54.1%)
Positive	265 (59.3%)	207 (35.6%)	472 (45.9%)
Progesterone receptor (PR)			
Negative	298 (66.7%)	66 (11.4%)	364 (35.4%)
Positive	149 (33.3%)	515 (88.6%)	664 (64.6%)
Her2.neu			
Negative	154 (34.5%)	330 (56.8%)	484 (47.1%)
Positive	293 (65.5%)	251 (43.2%)	544 (52.9%)
Initial grade			
1	12 (2.7%)	195 (33.6%)	207 (20.1%)
11 and 111	435 (97.3%)	386 (66.4%)	821 (79.9%)
Initial chemotherapy			
No	303 (67.8%)	351 (60.4%)	654 (63.6%)
Yes	144 (32.2%)	230 (39.6%)	374 (36.4%)
Initial radiotherapy			
No	31 (6.9%)	99 (17.0%)	130 (12.6%)
Yes	416 (93.1%)	482 (83.0%)	898 (87.4%)

395

Table 2: Multivariate analysis of parametric time to failure models of breast cancer patients

Variable	***Weibull***	***Exponential***	***Log-normal***	***Log-logistic***	***Rayleigh***
Age at diagnosis	1.037 (0.304)	1.066 (0.240)	1.058 (0.211)	1.074 (0.094)	1.024 (0.384)
Age at recurrence	0.969 (0.387)	0.945 (0.307)	0.952 (0.278)	0.938 (0.135)	0.981 (0.481)
Estrogen receptor (ER)	0.995 (0.935)	0.922 (0.421)	0.935 (0.383)	0.933 (0.323)	1.026 (0.612)
Progesterone receptor (PR)	2.964 ($<2e^{-16}$)	4.357 ($<2e^{-16}$)	2.955 ($<2e^{-16}$)	2.912 ($<2e^{-16}$)	2.461 ($<2e^{-16}$)
Her2.neu	0.734 ($4.2e^{-06}$)	0.680 (0.000)	0.718 ($3.3e^{-05}$)	0.723 ($1e^{-05}$)	0.757 ($5e^{-08}$)
Initial grade	0.196 (3.6^{-16})	0.099 ($4.6e^{-15}$)	0.231 ($<2e^{-16}$)	0.215 ($<2e^{-16}$)	0.270 ($<2e^{-16}$)
Chemotherapy	1.181 (0.015)	1.215 (0.062)	1.243 (0.007)	1.209 (0.011)	1.165 (0.004)
Radiotherapy	0.686 (0.002)	0.609 (0.008)	0.652 (0.001)	0.629 (0.000)	0.731 (0.001)

Values present the acceleration factor=exp(value) and in brackets the p-value.

398 **Table 3:** Overview of the Akaike Information Criteria (AIC) scores

Parametric models	***AIC***
Weibull	7347.873
Exponential	7441.794
Log-normal	7410.574
Log-logistic	7361.512
Rayleigh	7400.741

399

400 **Table 4:** Final Weibull model after backward step-wise selection

	Estimate	***Standard error***	***z***	***p***
Intercept	8.9424	0.2330	38.37	$<2e^{-16}$
PR	1.1065	0.0736	15.03	$<2e^{-16}$
Her2.neu	-0.3158	0.0665	-4.75	$2e^{-06}$
Grade 11 and 111	-1.6390	0.1993	-8.22	$<2e^{-16}$
Chemotherapy	0.1904	0.0668	2.85	0.004
Radiotherapy	-0.3897	0.1232	-3.16	0.002
Log(scale)	-0.4224	0.0390	-10.83	$<2e^{-16}$

401

Figures

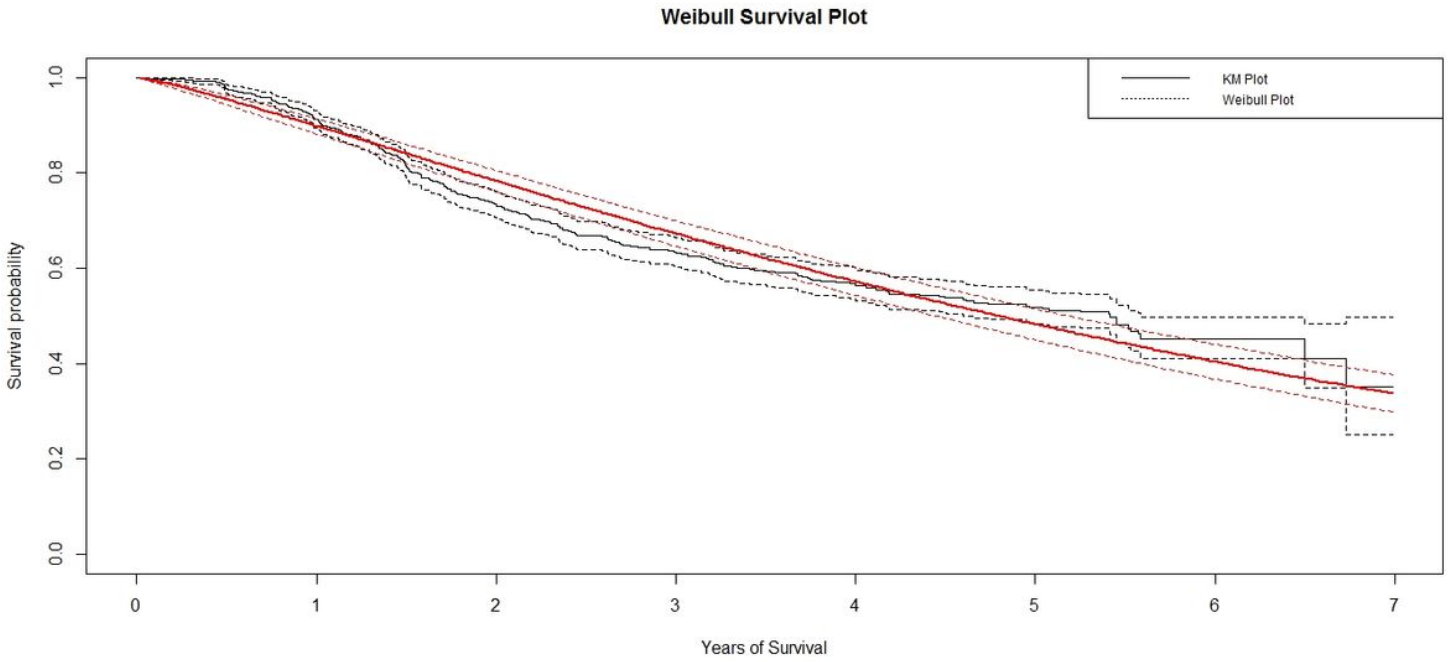

Figure 1

Weibull survival plot