RESEARCH ARTICLE

Survival models for right censored breast cancer data: theory, application and comparison [version 1; peer review: 2 not approved]

Madiha Liaqat¹, Shahid Kamal¹, Florian Fischer^{id}2,3, Waqas Fazil⁴

¹College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan

²Institute of Public Health, Charité - Universitätsmedizin Berlin, Berlin, Germany

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

⁴Institute of Nuclear Medicine & Oncology Lahore, Lahore, Pakistan

v1 First published: 13 Oct 2021, 10:1042
<https://doi.org/10.12688/f1000research.73507.1>
 Latest published: 13 Oct 2021, 10:1042
<https://doi.org/10.12688/f1000research.73507.1>

Abstract

Background: Censoring frequently occurs in disease data analysis, which is a key characteristic of time to failure modeling. Typically, time to failure studies are conducted through non-parametric and semi-parametric modelling techniques. Parametric models provide more efficient estimates, but are seldomly used, because of some of the limitations and assumptions which need to be fulfilled to apply them. The aim of this study is to illustrate the theoretical and application limitations and performance of different flexible and standard parametric models to evaluate the prognostic value for mortality risk of breast cancer after recurrence among women.

Methods: This article describes the theoretical properties of flexible parametric models and compares their performances to standard parametric models, by studying mortality in women diagnosed with breast cancer. We describe how time to failure data may be analyzed with nonlinear flexible models. In this regard, we apply fractional polynomials, spline models, piecewise exponential models, and piecewise exponential additive mixed models. We also illustrate properties of standard parametric models. All analyses have been conducted with multiple covariates to identify significant predictors. Information criteria have been used to evaluate performances of models.

Results: Fractional polynomial and spline-based generalized additive models work well in capturing local fluctuations. Parameter estimation with a piecewise exponential additive mixed model (PAMM) as an extension of the piecewise exponential modelling (PEM) approach automatically penalizes model complexity, which is very helpful to avoid over fitting.

Conclusions: Flexible parametric time to failure models are more efficient than standard parametric time to failure models. By

Open Peer Review

Approval Status **XX**

	1	2
version 1 13 Oct 2021	X view	X view
1. Benjamin Kearns ^{id} , The University of Sheffield, Sheffield, UK		
2. Federico Ambrogi ^{id} , University of Milan, Milan, Italy		
Any reports and responses or comments on the article can be found at the end of the article.		

incorporating time dependent covariates, PAMM is a good approach to perform in-depth studies of predictors over different finite intervals of follow-up time. Until now, this approach is rarely used in time to failure right censored studies.

Keywords

censoring, time to failure analysis, non-proportionality, splines, piecewise exponential models, piecewise exponential additive models, accelerated failure time, oncology

This article is included in the **Oncology** gateway.

Corresponding author: Florian Fischer (florian.fischer1@charite.de)

Author roles: **Liaqat M:** Conceptualization, Formal Analysis, Writing – Original Draft Preparation; **Kamal S:** Supervision, Writing – Review & Editing; **Fischer F:** Supervision, Writing – Review & Editing; **Fazil W:** Data Curation, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The work was supported by the Higher Education Commission Pakistan under grant No. 46-2SS2- 123 awarded to Madiha Liaqat. We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Fund of Charité – Universitätsmedizin Berlin.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Liaqat M *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Liaqat M, Kamal S, Fischer F and Fazil W. **Survival models for right censored breast cancer data: theory, application and comparison [version 1; peer review: 2 not approved]** F1000Research 2021, 10:1042 <https://doi.org/10.12688/f1000research.73507.1>

First published: 13 Oct 2021, 10:1042 <https://doi.org/10.12688/f1000research.73507.1>

Background

Cancer causes a large disease burden worldwide, among which breast cancer is the most frequent cause of cancer deaths in women. Pakistan, being a lower-middle-income country, has a greater number of breast cancer patients compared to its neighboring countries. It is the country with the **highest age-standardized death rate in 2019 globally**. Risk of death increases after early breast cancer recurrence in the first three to five years of primary treatment. Time after recurrence to death is analyzed through time to failure techniques, incorporating recorded prognostic factors before recurrence such as age, tumor grade, molecular subtype and treatment.¹ In previous research, age has not been proven to be a significant influence on breast cancer deaths.^{2,3} To further explore its role, age at diagnosis and age at recurrence are included in this study with other covariates.

Time to failure data have incomplete information about exact event occurrence time, which is known as censoring. Three common types of censoring are encountered in time to failure studies: right, left and interval censoring. The most common is right censoring, which is classified into three types: fixed type 1, type 11 and random type 1. In fixed type 1, right censoring occurs for all understudy subjects, who do not observe the event of interest during the predefined study time. Type 11 censoring is named for all subjects who do not observe a specific event after a pre-specified number of events have occurred. In random type 1 right censoring, censored subjects have different censoring times, as not all have same entry time into the study.⁴ Non-parametric, semi-parametric and parametric modelling techniques are amenable to analyze such types of time to failure disease studies.⁵

Kaplan-Meier (KM) is the simplest method, used to estimate survival function by a non-parametric maximum likelihood estimator (NPML), which has the limitation of studying only one factor at a time. Therefore, it is not suitable for multivariate studies.⁶ The Cox proportional hazard (PH) models, a semi-parametric approach, does not assume the shape of the baseline hazard function, so distributions of regression parameters' outcomes remain unknown.⁷ Cox PH models incorporate multivariate predictors by holding the PH assumption, which assumes a fixed proportion of hazard for individuals. In case of right censoring where upper bounds of event occurrences are not specified, regression parameters are estimated through dividing the likelihood function of the PH model into two parts: one comprises of the baseline hazard and unknown parameters, while the other has only unknown parameters to be estimated, which is called partial-likelihood. Breslow⁸ and Efron⁹ introduced approximations in partial-likelihood to handle ordered ties in uncensored event times, while exact and discrete methods are also available, in which non-ordered tied survival times are applied through a partial likelihood approach.⁵

Validity of PH assumption can be checked through a standard global test suggested by Grambsch and Therneau.¹⁰ Furthermore, graphical ways of plotting residuals versus predictors are also discussed in their research.¹⁰ In case of non-proportionality, extended Cox PH models can apply, which account for the effects of time varying predictors on survival times.¹¹ Spline-based methods are a good choice to estimate effects of unknown nonlinear predictors on continuous response through penalized partial likelihood, they also explore the functional form of non-proportional predictors.¹² Piecewise models are a good choice for long length follow-up studies, where predictors' effects are checked at different finite time intervals to obtain in-depth information about disease progression.¹³

Parametric models rely on a fully maximum likelihood approach, parametric estimates are more efficient and precise if conducted through correctly specified forms. In parametric modelling, time to failure is assumed to follow any distribution, such as exponential, Weibull, gamma, generalized gamma, log-normal, log-logistic, Gompertz and Generalized F.¹⁴ By building a linear relationship between the logarithm of failure time and predictors, data can be analyzed through the accelerated failure time (AFT) model. In AFT models one-unit changes in predictors explain a proportional change in survival time, as illustrated by Lee and Go,¹⁵ while in PH parametric form assumes a proportional change in hazard due to a one-unit change in predictors.

The aim of this paper is to review and apply the above stated modelling techniques to time to failure data, and to evaluate their performances through statistical measures, to investigate the best fitted one for right censored data, while fulfilling limitations and assumptions.

Methods

Study design

Our data consists of 1,028 women diagnosed with breast cancer in Lahore, Pakistan. All women observed recurrence between February 2011 and February 2018 after initial treatment. They were treated at the same hospital (Institute of Nuclear Medicine & Oncology Lahore, Pakistan). The primary endpoint of this study is death due to breast cancer. Exclusion criteria were: incomplete or missing information, women diagnosed with another disease or another cancer before breast cancer, and bilateral carcinomas. Women who were still alive (survived) at the end of the study, or died due

to another reason than breast cancer, are considered right-censored. Age at diagnosis, age at recurrence, estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (Her2), tumor grade, radiotherapy and chemotherapy are the predictors included in this study, all were chosen with the help of clinicians and oncologists.

In the data, age at diagnosis and age at recurrence (in years) are continuous variables, estrogen receptor, progesterone receptor and Her2 are represented by binary variables (0 = Negative, 1 = Positive), tumor grade is represented categorically (*I, II, III*). In addition, chemotherapy and radiotherapy are indicated by dummy variables (0 = No, 1 = Yes), here 0 indicated the patients who did not receive the treatment, while 1 = Yes meant they received treatment. Survival time was considered from recurrence to death or drop out, and censoring status was coded with 0 for censored, and 1 for death due to breast cancer. Along with the aim of this study, which is the comparison of different parametric models, it is also a major interest to find out how the two treatments (radiotherapy and chemotherapy), in combination with the other predictors, affect the survival of breast cancer patients after recurrence.

Proportional hazards have been checked by scaled Schoenfeld residuals (Extended data). Furthermore, the statistical tests revealed that not all covariates are statistically significant ($p < 0.05$), but the global test is statistically significant. Therefore, we can assume that the proportional hazard assumption holds.

Regression models

In this work, right censored time-to-event data are considered. Survival time is denoted by T and censoring time by C_j , where $j = 1, 2, 3, \dots, m$ are women diagnosed with invasive breast cancer who observed first recurrence after primary treatment. The j^{th} event time is defined by $t_j = \min(T_j, C_j)$, while time for censored and uncensored events is denoted by $\delta_j = I(T_j \leq C_j)$. The general relationship form, between T and X is given as

$$T = f(X) + \varepsilon, \quad (1)$$

where $X = x_1, x_2, x_3, \dots, x_p$, is the vector of predictors which may have an impact on time to failure, and f is the unknown functional form, which may be linear or non-linear. Practically, a 100% exact true relationship is not possible. To cover up uncontrolled chances of error, ε is also included in the model. $S(t) = P(T > t)$ is the survival function, which represents the probability of a woman survivor up to a time point, and the hazard rate is written as

$$\pi(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t) | (T \geq t)}{\Delta t} \quad (2)$$

The hazard rate represents a probability of an instantaneous failure per unit of time given that an individual patient has survival after time t .⁵ The main objective of time to failure studies is to estimate the hazard function accurately. For this purpose, different modelling approaches are applied.

In the multivariate approach, the semi-parametric Cox PH model is most popular for the analysis of time to failure data. The general form is written as

$$\pi(t|x_j) = \pi_0(t) \exp(x_j^T \alpha), \quad j = 1, 2, 3, \dots, m \quad (3)$$

where, m is the number of patients under study, and $x_j = (x_{j,1}, x_{j,2}, x_{j,3}, \dots, x_{j,p})^T$ is a row vector of p predictors for subject j , while α is the vector of regression coefficients.⁷ The partial likelihood method employed for estimating unknown parameters is suggested by Cox.¹⁶ In time to failure analysis, continuous predictors are often categorized, which has disadvantage of information loss within categories. Useful available statistical methods of handling continuous predictors are fractional polynomials (FPs) and restricted cubic splines (RCS).

FPs provide flexible parameterization of continuous predictors, were first used for modelling families of curves by Royston and Altman,¹⁷ with polynomial of degree m , it can be written as

$$FP_m(X) = \alpha_0 + \sum_{j=1}^m \alpha_j f_j(x_1) \quad (4)$$

m is a positive integer and $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_m$ are regression parameters, while $f_j(x_1)$ is divided into two parts: one consists of $x_1^{p_j}$ when $p_j \neq 0$, and another of $\ln(x_1)$ for $p_j = 0$.¹⁷ FP models have a wide variety of shapes based on different transformations. The main issue with fractional polynomials is in choosing a suitable power for polynomials, as this has a

direct positive relationship with flexibility. To increase flexibility, a greater power can be used, but with the major threat of non-locality. That means that the fitted function at a given point of x_0 depends on data points which are very far from that reference point.¹⁸

Spline regression is an improved technique, used to overcome the non-locality of fractional polynomials. In spline models, the dataset is divided into multiple parts and these parts are joined with knots. In time to failure regression analysis, spline modelling is extensively used to smooth non-linear effects of continuous predictors. Spline $f(X)$ has a smooth function. The major problem occurs in choosing the number of knots, as no hard and fast rule is available to apply the suitable number of knots.¹⁹

Under RCS, the best way of choosing the number of knots is using the quotient of the difference between the largest and average uncensored log survival time and the largest and smallest uncensored log survival time.²⁰ Royston suggested another method in this respect; according to him, a good way to choose the suitable number of knots is to randomly apply different number of knots every time, and select the best model with measure of information criteria.²¹ Flexible parametric models have scaling for proportional hazards or proportional odds, which are usually based on transformation of survival function by a link function

$$\eta[S(t; x_j)] = \eta[S_0(t)] + \alpha^T x_j \quad (5)$$

$S_0(t)$ is the baseline survival function, α is the vector of unknown parameters for predictors x . Piecewise exponential models (PEMs) are also a reasonable approach to estimate hazard ratios more accurately.²¹ Under the PEM modelling approach, follow-up time is divided into i intervals, by assuming a constant baseline hazard in each interval, so that $\pi_0(t) = \pi_i$ simplifies to,

$$\pi_j(t|x_j) = \pi_i \exp(x_j^T \alpha) \quad (6)$$

For the cut points of follow-up times, minimum to maximum time is divided into finite intervals, $0 = n_0 < n_1 < n_2 < n_3 < \dots < n_i = t_{\max}$. Here, n_0 to n_i are time intervals, and the hazard rate of exponential distribution is constant over time intervals. Censoring and all unique event times can be used as time interval cut points, but no hard and fast rule is available to choose cut points, which is a point to ponder; too small or too large cut points may cause under- or overfitting.

One approach to deal with this cut point problem is an extension of PEM, in which a large number of cut points are used and the hazard is estimated semi-parametrically. This is called the Piecewise exponential additive mixed model (PAMM), in which a hazard is modeled through a smooth nonlinear function. In PAMM, predictors contribute to the hazard additively, imposing a quadratic penalty on the basis coefficients:

$$\pi_j(t|x_j) = \exp\left(f_0(t_j) + \sum_{n=1}^p f_n(x_{j,n}, t_j)\right), \quad \forall t \in (n_{i-1}, n_i) \quad (7)$$

where, $f_0(t_j)$ denotes log baseline hazard rate, and $f_n(x_{j,n}, t_j)$ represented effects of smooth nonlinear constant predictors, while t_j is finite time cut point. PAMM is an extension of PEM, in which modeling is done by using baseline hazard as a spline basis function, hazard is constant across intervals, through penalization over fitting is avoided even with large number of time intervals.²²

In fully parametric PH modelling, baseline hazard function is assumed to follow a specific distribution and coefficients are estimated via maximum likelihood. A number of different parametric PH models are derived by applying distributions, such as exponential, Weibull, and Gompertz.

An alternative to parametric PH is AFT models, the corresponding log-linear form of the AFT model with respect to time is given as

$$\log T_j = \alpha_0 + \sum_{j=1}^p x'_j \alpha_j + W \varepsilon_j \quad (8)$$

where, α_j is the vector of coefficients of unknown parameters, W is the scale parameter, and ε is the random error term.²³ AFT models measure the direct effect of predictors on the survival time rather than the hazard. ε_j , as a random variable, assumes different distributions for survival time T , such as exponential, Weibull, gamma, generalized gamma, log-normal, log-logistic and generalized F.²⁴

The likelihood estimates are maximized using the Newton Raphson procedure,¹⁵ which may be time consuming and tricky without computer programming. The freely available **R software** is used to implement all modelling techniques.

Measures of models fitting

Comparison of fitting models is done via measures of fit, which describe accuracy of fitted models for a given data set, usually called goodness of fit measures. Model fitting accuracy has nothing to do with the predictive ability for external data prediction. The Akaike Information Criterion (*AIC*)²⁵ and the Bayesian Information Criterion (*BIC*)²⁶ are two of the most common measures which are used to compare models' performances. For the PH model *AIC* and *BIC* are based on log-partial likelihood

$$AIC = -2(\text{Log - partial likelihood}) + 2(df) \quad (9)$$

$$BIC = -2(\text{Log - partial likelihood}) + n(df) \quad (10)$$

Table 1. Characteristics of multivariate covariates of breast cancer time to failure understudy data.

Covariates	Uncensored (n = 447)	Censored (n = 581)	Total (n = 1,028)
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 (in 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)
Human epidermal growth factor receptor 2 (Her2)			
Negative	154 (34.5)	330 (56.8)	484 (47.1)
Positive	293 (65.5)	251 (43.2)	544 (52.9)
Initial grade			
I	12 (2.7)	195 (35.5)	207 (20.1)
II	121 (27.1)	255 (43.8)	376 (36.6)
III	314 (70.2)	131 (22.5)	445 (43.2)
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)

Uncensored: Deaths due to breast cancer, Censored: Deaths due to other reasons or still alive patients.

where, df represents degrees of freedom of the fit, and n is the total number of observations in the data. The minimum value of both AIC and BIC is considered a good one. Basically, the AIC criterion is used as a penalized function, as if one adds a variable, sampling variability also increases. While BIC imposes stronger penalty in the inclusion of additional covariates to the model. Hurvich *et al.*²⁷ suggested a modified AIC , in which $n \times (df + 1) / (n - (df + 2))$ is used as a penalty term. A corrected version of BIC proposed by Volinsky and Raftery²⁸ can also be used, which replaced n with uncensored observations. Corrected AIC ($AICc$) and BIC ($BICc$) have written forms as

$$AICc = -2(\text{Log} - \text{partial likelihood}) + \frac{2(n)(df + 1)}{n - (df + 2)} \quad (11)$$

$$BICc = -2(\text{Log} - \text{partial likelihood}) + n_{\text{uncensored}}(df) \quad (12)$$

Ethical 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. The study was critically cleared by the Advanced Studies and Review Board of Punjab University. The letter of support written by the departmental head was submitted to the selected hospital. Prior to data collection, written consent was obtained from the head of oncology department and confidentiality was maintained by coding from data collection to analysis.

Results

In the present study, women's age was collected for: diagnosis and recurrence time. 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 cases were ER-negative. The majority of patients were PR-positive (64.6%) and had a positive human epidermal growth factor receptor 2 (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 included in the study. As shown in Table 1, 78.5% of deaths occurred due to breast cancer within three years after recurrence, while 20.8% between 3 to 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% human epidermal growth factor receptor 2-positive. Breast cancer death was positively associated with higher tumor grade (11 and 111; 97.3%) and no chemotherapy (67.8%).

Table 2. Log-likelihood and information criteria for standard parametric accelerated failure time and flexible parametric models.

	–2 log likelihood	Parameters	AIC	AICc	BIC	BICc
Exponential	7375.2	10	7395.3	7397.4	17655.2	11845.2
Weibull	7247.6	11	7269.5	7271.9	18555.6	12164.6
Gamma	7248.2	11	7270.2	7272.5	18556.2	12165.2
Generalized Gamma	7245.8	12	7269.8	7272.1	19581.8	12609.8
Log-normal	7327.8	11	7349.8	7352.1	18635.8	12244.8
Log-logistic	7265.6	11	7287.5	7289.9	18573.6	12182.6
Generalized F	7245.6	13	7271.7	7274.0	20609.6	13056.6
FP	5155.5	6	5167.5	5163.6	11323.5	7837.5
GAM	5153.5	9.06	5171.3	5173.8	14467.2	9203.3
R-P odd	7248.9	15	7278.9	7281.4	22668.9	13953.9
R-P hazard	7214.0	15	7244.0	7245.5	22634.0	13919.0
R-P (G G) odd	7160.5	26	7212.5	7216.0	33888.5	18782.5
R-P (G G) hazard	7159.3	26	7211.3	7214.8	33887.3	18781.3

AIC: Akaike's information criterion; FP: fractional polynomial; GAM: generalized additive model; RP: Royston-Parmar models.

Table 2 presents information measure results. Low values of AIC , $AICc$, BIC and $BICc$ are considered good; if a model's fitting values for AIC , $AICc$, BIC and $BICc$ are smaller than others, that model is considered a good fitted one. To make results less lengthy and meaningful, we only discuss here AIC and $AICc$ values of first three good fitted accelerated failure time distributional models. From the fully parametric models, Weibull is the best fitted one ($AIC = 7269.5$, $AICc = 7271.9$) among others, generalized gamma ($AIC = 7269.8$, $AICc = 7272.1$), gamma ($AIC = 7270.2$, $AICc = 7272.5$), and Generalized F ($AIC = 7271.7$, $AICc = 7274.0$) come next in terms of preferences, respectively. We also presented BIC and $BICc$ in **Table 2**.

Multivariate fractional polynomial (MFP) is the best fitted model incorporating time dependent covariates ($AIC = 5167.5$, $AICc = 5163.6$), while the generalized additive model (GAM) ($AIC = 5171.3$, $AICc = 5173.8$) is also a good choice for analyzing non-linear continuous predictors in a multivariate setting, having the advantage of small numbers of parameters in non-integer, which is due to shrinkage during parameter estimation. Royston and Parmar's flexible parametric models

Figure 1. Observed and modeled hazards.

Figure 2. Nelsom Aalen, PAM, PAMM cumulative hazard graph.

have been applied with different scales: we considered flexible parametric models on hazard and odd scales, by including time-dependent effects for the age at diagnosis and age at recurrence covariates. Hazard generalized gamma ($AIC = 7211.3$, $AICc = 7214.8$) outperformed odd generalized gamma ($AIC = 7212.5$, $AICc = 7216.0$). Although the subjective approach of knot selection may be criticized, sensitivity analyses studies showed insignificant differences in results while changing positions of knots.^{29–31}

Figure 1 shows cumulative hazard graphs for all parametric and flexible parametric models with Nelson Aalen cumulative hazard as a reference. The Nelson Aalen estimator is represented by a step function, which starts at zero. It provides an estimate of the expected number of deaths observed for a given amount of time. Visually, all models provide fitting accuracy for the right censored breast cancer failure time understudy data, with some slight variations existing to capture the fluctuations. The wider lines show a greater confidence interval, which is indicative of a poor fit, while narrower lines show good model fitting.³²

Time dependent covariates, age at diagnosis and age at recurrence are modeled using 4 degrees of freedom for splines. PEM and PAMM are applied to the understudy data to get baseline hazard estimates, where finite time intervals are considered as factors to maintain constant hazard for each interval. The age at diagnosis and recurrence are estimated using P-splines with the same 4 degrees of freedom. PEM and PAMM results are compared with Nelson Aalen estimator, graphical displays showed close agreement of good model fitting in Figure 2.

Discussion

In this paper, different parametric models are compared in terms of theoretical aspects and application. Our findings suggested that progesterone receptor negative, human epidermal growth factor receptor 2 positive, higher tumor grade, and no chemotherapy increase the risk of death after recurrence. The most surprising result is regarding 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.³³

We applied distributional parametric models which are known as standard parametric models, with a full maximum likelihood estimator to estimate unknown parameters. AFT models make practical sense to study the influence of covariates, which may accelerate breast cancer mortality. 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.

The exponential distribution having one rate parameter is often used in experiments to account for the amount of time until an event occurs. The Weibull distribution is a special case of the exponential distribution with shape and scale parameters. It provides a better fit than exponential with one extra degree of freedom. The same is true for gamma distribution, which has two parameters and has close results to Weibull. Generalized gamma has mean, location and scale

parameters. Log-normal is a probability distribution, with a normally distributed logarithm. It is widely used in lifetime data analysis. The two parameters of mean and standard deviation have a more stable behavior than log-logistic distribution. Generalized F distribution is a good alternative to generalized gamma with one extra parameter. From the interpretation point of view, the AFT model's results are easy to interpret and help clinicians to make wise decisions related to the patients' conditions.

Flexible parametric models have advantage of using restricted cubic splines, which incorporates time dependent effects of predictors on the log hazard and reduces the bias of non-proportionality. The Royston and Parmar generalized gamma flexible parametric model under hazard scale outperformed the odd scale model. Of course, one should not ignore the threat of overfitting, by including greater number of internal knots. The functional polynomial model has the advantage of only considering significant factors, so it gives better results than other spline-based models. GAM under spline basis function has the potential to provide a better fit of data than generalized linear flexible parametric models.^{34,35}

The main strength of this study is that we described and applied different time to failure models, to right censored breast cancer data. In a piecewise exponential model the baseline hazard is modeled by step function with different intervals, estimation is done by including dummy variables for each interval. The major disadvantage of this technique is that data becomes too long, and parameter estimation becomes unstable. The piecewise additive mixed model overcomes this drawback. By adding a large number of basis functions and using P-splines between neighboring basis coefficients, parameters are estimated through restricted maximum likelihood (REML).¹³

Limitations

There are several limitations to our study. First, the use of a single case study may be viewed as a limitation. However, a simulation study can also be designed to validate results. Second, the model comparisons are based on within sample information measures (*AIC*, *AICc*, *BIC*, *BICc*), while predictive performances of models can also be checked via different measures. Third, sensitivity analyses of choosing different numbers of knots in spline-based models can be performed to make firm conclusions.

Conclusion

Flexible parametric modelling of the hazard function is more efficient than standard parametric models, incorporating the complex patterns of the observed failure data. Generalized additive models provide more accurate estimates under spline-basis function, with time dependent covariates. For long follow-up studies and multiple time dependent covariates, which may have effects on hazard, penalized models are more suitable.

Data availability

Underlying data

Data is available from the corresponding author, Dr. Florian Fischer (florian.fischer1@charite.de), upon reasonable request. Data can be used for research purposes, but cannot be published because it is taken from a hospital.

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, for providing full support to conduct this research.

References

- McPherson K, Steel CM, Dixon JM: **ABC of breast diseases. breast cancer-epidemiology, risk factors, and genetics.** *BMJ*. 2000; **321**: 624–628.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Barchielli A, Balzi D: **Age at diagnosis, extent of disease and breast cancer survival: a population-based study in Florence, Italy.** *Tumori*. 2000; **86**: 119–123.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Crowe JP Jr, Gordon NH, Shenk RR, et al.: **Age does not predict breast cancer outcome.** *Arch. Surg.* 1994; **129**: 483–487.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lagakos SW: **General right censoring and its impact on the analysis of survival data.** *Biometrics*. 1979; **35**: 139–156.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lee E: *Statistical Method for Survival Data Analysis*. New York: Wiley; 1992.
- Kaplan EL, Meier P: **Nonparametric estimation from incomplete observations.** *J. Am. Stat. Assoc.* 1958; **53**: 457–481.
[Publisher Full Text](#)
- Cox DR: **Regression models and life-tables.** *J. Royal Statistical Society. Series B (Methodological)*. 1972; **34**: 187–202.
[Publisher Full Text](#)

8. Breslow N: **Covariance analysis of survival data under the proportional hazards model.** *Int. Stat. Rev.* 1974; **43**: 45–54.
[Publisher Full Text](#)
9. Efron B: **The efficiency of Cox's likelihood function for censored data.** *J. Am. Stat. Assoc.* 1977; **72**: 557–565.
[Publisher Full Text](#)
10. Grambsch PM, Therneau TM: **Proportional hazards tests in diagnostics based on weighted residuals.** *Biometrika.* 1994; **81**: 515–526.
[Publisher Full Text](#)
11. Cox DR, Oakes D: *Analysis of Survival Data.* New York: Chapman & Hall; 1984.
12. Fahrmeir L: **Dynamic modelling and penalized likelihood estimation for discrete time survival data.** *Biometrika.* 1994; **81**: 317–330.
[Publisher Full Text](#)
13. Bender A, Fabian S, Wolfgang H, *et al.*: **Penalized estimation of complex, non-linear exposure-lag-response associations.** *Biostatistics.* 2018; **20**: 315–331.
[Publisher Full Text](#)
14. Crowther MJ, Lambert PC: **A general framework for parametric survival analysis.** *Stat. Med.* 2014; **33**: 5280–5297.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Lee ET, Go OT: **Survival analysis in public health research.** *Annu. Rev. Public Health.* 1997; **18**: 105–134.
[Publisher Full Text](#)
16. Cox DR: **Partial likelihood.** *Biometrika.* 1975; **62**: 269–276.
[Publisher Full Text](#)
17. Royston P, Altman DG: **Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling.** *J. Royal Statistical Society. Series C (Applied Statistics).* 1994; **43**: 429–467.
[Publisher Full Text](#)
18. Royston P, Sauerbrei W: *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* Wiley; 2008.
[Publisher Full Text](#)
19. Durrleman S, Simon R: **Flexible regression-models with cubic-splines.** *Stat. Med.* 1989; **8**: 551–561.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Royston P: **Flexible alternatives to the Cox model, and more.** *The State Journal.* 2001; **1**: 1–28.
[Publisher Full Text](#)
21. Friedman M: **Piecewise exponential models for survival data with covariates.** *Ann. Stat.* 1982; **10**: 101–113.
[Publisher Full Text](#)
22. Bender A, Andreas G, Fabian S: **A generalized additive model approach to time-to-event analysis.** *Stat. Model.* 2018; **18**: 299–321.
[Publisher Full Text](#)
23. Wei LJ: **The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis.** *Stat. Med.* 1992; **11**: 1871–1879.
[PubMed Abstract](#) | [Publisher Full Text](#)
24. Viliandias B, Mikhail N: *Accelerated Life Models; Modeling and Statistical Analysis.* Chapman&Hall/CRC; 2002.
25. Akaike H: **A new look at the statistical model identification.** *IEEE Trans. Autom. Control.* 1974; **19**: 716–723.
[Publisher Full Text](#)
26. Schwarz G: **Estimating the dimension of a model.** *Ann. Stat.* 1978; **6**: 461–464.
[Publisher Full Text](#)
27. Hurvich CM, Tsai CL: **Regression and time series model selection in small samples.** *Biometrika.* 1989; **76**: 297–307.
[Publisher Full Text](#)
28. Volinsky CT, Raftery AE: **Bayesian information criterion for censored survival models.** *Biometrics.* 2000; **56**: 256–262.
[PubMed Abstract](#) | [Publisher Full Text](#)
29. Lambert PC, Dickman PW, Nelson CP, *et al.*: **Estimating the crude probability of death due to cancer and other causes using relative survival models.** *Stat. Med.* 2010; **29**: 885–895.
[PubMed Abstract](#) | [Publisher Full Text](#)
30. Royston P, Lambert PC: *Flexible parametric survival analysis using Stata: beyond the Cox model.* Stata Press Books; 2011.
31. Nelson CP, Lambert PC, Squire IB, *et al.*: **Flexible parametric models for relative survival, with application in coronary heart disease.** *Stat. Med.* 2007; **26**: 5486–5498.
[PubMed Abstract](#) | [Publisher Full Text](#)
32. Jackson CH: **flexsurv: a platform for parametric survival modeling in R.** *J. Stat. Softw.* 2016; **70**: 1–33.
33. Bhoo-Pathy N, Verkooijen HM, Wong FY, *et al.*: **Prognostic role of adjuvant radiotherapy in triple-negative breast cancer: a historical cohort study.** *Int. J. Cancer.* 2015; **137**: 2504–2512.
[PubMed Abstract](#) | [Publisher Full Text](#)
34. Wood SN: *Generalized Additive Models: An Introduction with R.* Boca Raton (FL): CRC Press; 2017.
[Publisher Full Text](#)
35. Hastie T, Tibshirani R: *Generalized Additive Models.* New York: Wiley Online Library; 1990.

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 27 September 2022

<https://doi.org/10.5256/f1000research.77162.r149019>

© 2022 Ambrogio F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Federico Ambrogio

Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy

After reading the title, I expected to read a tutorial or review paper for survival analysis for breast cancer patients. The paper is in part a review but I discovered that the focus is on parametric models and has some more general aspects scattered throughout the text.

The main critique has to do with the endpoint chosen by the authors. As death from breast cancer is used, some considerations about competing risks are necessary and nothing is said in the paper. The analysis performed using regression model is that of cause-specific hazards and generally must also take into account the other causes of death. There are plenty of tutorial papers on competing risks both in the methodological and applied medical literature. Moreover, the study population is described approximately. The type of recurrence, for example, is not specified: is it the same for all 1028 women? What is the time scale for the analysis? Is the time interval starting from the time of recurrence until eventual death? One option is to use age as time scale and consider late entry.

The Results section is also questionable, having to deal with time to event data. How was median survival time calculated? In general, we have to take into account censored observations. The cumulative incidence curve (accounting for competing events) must be used to calculate the median event time! The same applies to the percentages of women with the events within 3 years and so on (Results section): this must be calculated using cumulative incidence curves, not just calculating the percentages as censoring makes percentages meaningless.

General conclusions about models cannot be drawn from this study and the sentence about simulation is too vague to be of any utility. In my opinion, this could be a tutorial paper about regression models in survival analysis but the data used are really too complicated for a tutorial!

Specific comments:

- The explanation of the different censoring type is not of interest here. Instead, late entry and time scales are of interest considering the data.

- Multivariate must be changed to multivariable.
- KM can be used with more than one variable. Obviously if using many variables, an excessive stratification may prevent any meaningful result.
- The sentence is not clear "...so distributions of regression parameters' outcomes remain unknown." Regression parameters in the Cox model have clear statistical properties.
- What are "unknown nonlinear predictors" and how can splines model them?
- PH assumption "assumes a fixed proportion of hazard for individuals"? Must be better explained!
- What is the distinction between binary and dummy variables?
- Model (1) is the "general relationship form" or something very special?
- "...the hazard rate of exponential distribution is constant over time intervals", probably the hazard is constant in each interval.
- "...too small or too large cut points may cause under- or overfitting", probably the time intervals are too large or too small.
- "The likelihood estimates are maximized using the Newton Raphson procedure,¹⁵ which may be time consuming and tricky without computer programming." This is a sentence from the fifties...
- "...if a model's fitting values for AIC, AICc, BIC and BICc are smaller than others, that model is considered a good fitted one", it is considered better than the others...
- MFP and GAM information criteria are probably not on the same scale as the others and comparison cannot be direct.
- Explanation of the MFP is lacking...

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Multivariate analysis; Survival analysis; Study design; high-dimensional data.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 19 January 2022

<https://doi.org/10.5256/f1000research.77162.r118775>

© 2022 Kearns B. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Benjamin Kearns

School of Health and Related Research, The University of Sheffield, Sheffield, UK

This is an interesting manuscript, but it feels very unfocused. For example, the first paragraph mentions the importance of age, and evaluating the effect of treatment is also mentioned, along with discussion of PH and AFT. This suggests that identifying and quantifying treatment effects is the main objective. But there are no model-based estimates of covariate effects provided in the results section.

There is also an extensive discussion and comparison of different model types (which is stated to be the main focus). This includes non and semi parametric approaches, but they are not presented in the results section. Parametric models are included, but it is unclear how the choice between these should be made (noting that there is no way to compare information criteria for statistical significance), and it is unclear how generalisable the results of this study are beyond the case-study provided. The abstract concludes "PAMM is a good approach to perform in-depth studies of predictors over different finite intervals of follow-up time." This is not supported by the results presented (and this conclusion is missing from the main text).

There are some important omissions, such as technical details for the GAM, and model specifications that were used for analysis (such as which degrees of FPs, basis functions for GAMs) with justification.

There are some places where detailed information is provided which does not really contribute to the manuscript. Examples include a discussion of the types of right censoring in the introduction

(this distinction is never used elsewhere in the manuscript), and discussion of both non- and semi-parametric methods, when these do not appear in the results. It is also unclear why Figures 1 and 2 are separate (and graphical results for the RP GG models look wrong, they have the lowest IC of the models presented in Fig 1, but the worst visual fit).

Overall, the manuscript requires substantial additional restructuring to make it suitable for publication. I would recommend making the focus a tutorial-style paper to demonstrate how flexible models may be used to estimate time-varying treatment effects, and how this compares with the treatment-effects obtained from standard approaches. This could focus on the impact of treatment. To support this, the R code used should be made available (even if the data cannot be), to enhance reproducibility. The authors could also consider replicating the analysis on a publicly available dataset (see for example Kearns *et al.* 2019¹)

Some additional feedback is provided below:

- Background: first paragraph needs more references to support the statements made.
- Background: first paragraph needs more justification for why the role of age is being explored when it was previously not found to be significant. What makes the authors think they will find a different association?
- As noted, most of the flexible models (such as GAMs and FPs) were originally developed for non-survival data, so information is required as to how they can be applied to survival data.
- Methods, study design: it is unclear what the "Extended data" is.
- Methods: π_0 (for the Cox PH) needs defining.
- Use of information criteria will be limited as non- and semi-parametric models cannot then be compared. It is unclear why AIC(c) were used in preference to BIC(c) when presenting results. It is also unclear what would happen if the four IC measures gave conflicting results (which one would be used to select the best model)?
- Results, Table 2: it is unclear why IC are so much lower for FPs and GAMs - this suggests that the likelihood for these two models is defined differently.
- Discussion: the benefits of AFT models will only hold if the aft assumption holds this is an important caveat that should be mentioned. Also, as the PH assumption is earlier stated to hold for this analysis, the relevance of the discussion of AFT models is unclear.
- Discussion: "The Weibull distribution is a special case of the exponential distribution" - it is the other way around. This paragraph is on the whole too general.
- Discussion: "Flexible parametric models have advantage of using restricted cubic splines" - there are a large number of flexible models that use other basis functions (or alternative approaches to induce flexibility).

References

1. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A: Generalized Linear Models for Flexible Parametric Modeling of the Hazard Function. *Med Decis Making*. **39** (7): 867-878 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Methodological research; health economics, survival analysis, time-series analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research