

Research Paper



Evaluation of different survival analysis models for NKI breast cancer data

John Edmon Alejandro Ganas^{1*}, Peter John Berces Aranas²

^{1*}Graduate School, Polytechnic University of the Philippines-Manila, Philippines.

²School of Statistics, University of the Philippines-Diliman Philippines.

Article Info

Article History:

Received: 14 July 2023

Revised: 20 September 2023

Accepted: 28 September 2023

Published: 06 November 2023

Keywords:

Survival Analysis Models

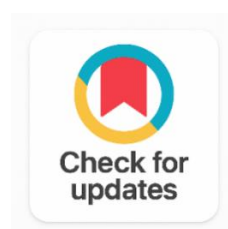
Concordance Index

Akaike Information Criterion (AIC)

Breast Cancer

Semiparametric Models

Parametric Models



ABSTRACT

Background: Survival analysis is pretty central in oncological research, but choosing the right model for a specific dataset is still a major, methodological problem. In this work we look at six survival analysis models on the NKI Breast Cancer dataset over an 18-year clinical trial period to see which one sort of performs best for survival prediction, yes.

Objective: The aim is to contrast semi-parametric versus parametric survival analysis approaches using Concordance Index (C-index) and Akaike Information Criterion (AIC), and then decide which one is most suitable for building a breast cancer survival prediction tool.

Methods: We used right-censoring to handle the cases where subjects were lost to follow-up. A proportional hazards test was applied to check whether Cox-based models make sense, and for the parametric side we ran a time-to-event distribution test to judge if those models are applicable. For the semi-parametric group, we evaluated three methods: Classical Cox, Cox-Lasso, and Cox-Ridge Regression. For the parametric group, we also tested three Accelerated Failure Time (AFT) models: Weibull AFT, Log-logistic AFT, and Log-Normal AFT.

Results: In the semi-parametric category, Cox-Ridge Regression came out on top with the highest C-index (0.7709) and the lowest AIC (752.6703). It performed better than Classical Cox and Cox-Lasso, in a fairly clear way. For the parametric models, Log-Normal AFT was best overall, showing a C-index of 0.780 and an AIC of 608.822, which beat both Weibull and Log-logistic AFT. When we compare across models overall, Log-Normal AFT stays the best-performing option.

Conclusions: The Log-Normal AFT option came out ahead of all the other semi parametric and parametric alternatives on both evaluation criteria, like clearly better. From these results it seems the survival times in the NKI Breast Cancer dataset match a log normal distribution fairly well, and so it is suggested that Log-Normal AFT be used as the starting point or basis for a breast cancer survival prediction model.

Corresponding Author:

John Edmon Alejandro Ganas

Graduate School, Polytechnic University of the Philippines-Manila, Philippines.

Email: jeaganas@pup.edu.ph

Copyright © 2023 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Over the course of advancement in the discipline of Statistics and its significant application to different facets of human endeavor, a lot of groundbreaking findings have been included in the current pool of humanity's collective knowledge. It has been evident that statistical methods are highly integrated in different aspects of healthcare sciences and medicine. According to the study made by [1], "results were summarized for statistical methods used in the literature, including descriptive and inferential statistics, modeling, advanced statistical techniques, and statistical software used. Approximately 81.9% of articles reported an observational study design and 93.1% of articles were substantively focused. Descriptive statistics in table or graphical form were reported in more than 95% of the articles, and statistical inference reported in more than 76% of the studies reviewed." These findings exhibit the substantial usage of various basic to advanced statistical methods to public healthcare alone. By considering other sub-disciplines under medical sciences, specifically in cancer research is the usage of predictive studies. According to [2], "Predictive studies address multiple predictors with combined effects on response to treatment and outcome prediction." Cancer research, which also includes Cancer Survival Analysis and prediction, has been according to [3], "the soundest tool to generate new knowledge."

Prediction of Cancer Survival is best modeled through Survival Analysis, or in other texts, Time-to-Event Analysis, is "widely used in clinical and epidemiological research" [4]. Survival Analysis exudes an efficient utilization of data from Cancer Survival since the variable for survival itself is time dependent.

The goal of this study is to evaluate Different Survival Analysis Models in terms of their predictive capabilities, accuracy in determining significant covariates within the data, as well as their respective results compared across standard indices specifically C-Index and AIC [5] which will lead to the selection of the ideal model as a benchmark for the construction of a Statistical Model for simulating Breast Cancer Survival [6].

1.1. Research Elaborations

This study will implement the usage of Comparative Method wherein the researcher aims to look for similarities and differences between objects of concern (Comparative Methods, para. 1-2), which in the context of this study are the 6 different Survival Analysis Methods which will all be applied to a high-dimensional clinical dataset for Breast Cancer Survival.

This study will follow the same modeling scheme done by [7] on comparing and evaluating different Survival Analyses as a basis for a Machine Learning Algorithm for a high-dimensional dementia clinical data. As for this study, the concentration will be on 6 different Survival Analysis Models which will, in the similar manner of how the original researcher's approach, are to be evaluated and selected based on their respective Partial AIC and Concordance Index which will be used as the most ideal Survival Prediction Model for the clinical dataset used in this study.

The heterogeneous clinical data used for this study was sourced out from data. World's Netherlands Cancer Institute (NKI) Breast Cancer metadata which contains the 272 patients information, types of treatment they underwent, survival times and event of death (coded as 0=Censored and 1=Dead).

77 out of the 272 (28%) patients died while the remaining either survived or lost follow-up at the end of the 18.34-year clinical trial.

Before the data is fitted in the Cox Models, the assumption of Proportional Hazards must be satisfied first. This means that ratio of hazards for any observed cases must remain constant until the end of the clinical trial/observation period [8]. As for the AFT models, the assumption that must be satisfied should be each of the covariates' effect with respect to the survival time must act multiplicatively (proportionally) [9]. This suggests that AFT models work almost the same as those conventional linear models wherein the logarithm of the Survival Times act as the response variable [10]. In other words, if the covariates exhibit a deceleration/acceleration on the response variable Survival Time, no strict assumptions are needed. The data then run through the 6 different Survival Analysis Models namely: Classical Cox Regression, Lasso Cox, Ridge-Cox, Weibull AFT model, Log-Logistic AFT model, and Log-Normal AFT model. From here, the selection of an ideal Survival Analysis Model will be used as a basis for construction of a Machine Learning Model will take place based on the survival model with the HIGHEST Concordance Index and Lowest AIC [11].

The following functions are necessary in the study of survival analysis involving CENSORED (subjects whose status are unknown until the end of the observation period/trial or data gathering) data points.

2. RESULTS AND DISCUSSION

To verify if the implementation of Cox Regression and its variants would work on the data at hand, a Proportional Hazards Test was performed, work on the data at hand, a Proportional Hazards Test was performed.

Table 1. Test of Proportional Hazards

Covariates	Classical Cox	Cox-LASSO	Cox-Ridge
Age	0.46	0.37	0.46
Amputation	0.06	0.06	0.06
Angioinv	0.1	0.04	0.09
Chemo	0.24	0.25	0.23
diam (mm)	0.29	0.34	0.3
esr1	2.93	2.86	2.93
Grade	1.56	1.57	1.54
Histtype	1.72	1.74	1.73
Hormonal	0.19	0	0.18
Lymphinfil	0.17	0.16	0.16
Posnodes	0.47	0.4	0.46
*tested at 99% level of confidence			

The Null Hypothesis for this test states that Proportional Hazards is satisfied at a 99% level of confidence. This also indicates that a 1% likelihood that at least one of these covariates would not follow the Assumption of Proportional Hazards. Since all associated p-values for each test statistic for every covariate flagged as not significant, it has been verified that all these covariates assume Proportional Hazards, which then qualifies the usage of the Cox Models stated (Classical Cox, Cox-Ridge, and Cox-LASSO).

Table 2. Summary Table for Cox Models

Covariates	Classical Cox	Cox-LASSO	Cox-Ridge
Age	-0.0635*	-0.0623*	-0.063*
Chemo	-0.2852	-0.2742	-0.2862

Hormonal	0.0141	0.0001	0.0109
Amputation	0.1989	0.1936	0.1974
Histtype	0.4225	0.4138	0.4192
diam (mm)	0.0203	0.0201	0.0204
Posnodes	0.0441	0.0427	0.0444
Grade	0.8517*	0.8438*	0.8454*
Angioinv	0.2483	0.2442	0.2476
Lymphinfil	-0.6644*	-0.6518*	-0.6579*
esr1	-1.1143*	-1.1026*	-1.1085*
*tested at 99% level of confidence			

For comprehensive interpretation of all Cox Regression coefficients relative to their corresponding Hazards, positive coefficients, and Hazard Ratio > 1 are indicators of bad prognosis (decreases survival and increases risk) whereas negative coefficients and Hazard Ratio < 1 indicates a “protective” effect (increases survival and decreases risk) to the variable it is associated.

At a 99% level of confidence, it has been observed that across the 3 models, age, cancer grade, lymphinfil and estrogen 1 receptors are consistent significant predictors in this model. Moreover it can also be seen that age, extent of lymphocytic infiltration, and estrogen 1 receptors carry negative coefficients, suggesting that these variables are associated to an increase in the survival of the subjects. As for the cancer grade, it is expected that positive increment (increase in the grade) is attributed to the decline of subject survival. It is also worth noting that even though treatments (chemotherapy, hormonal treatment, and breast amputation) did not flag statistical significance, only chemotherapy is associated to an increased survival. Overall, regardless of which model is to be considered, across all 3 Cox variants, combination of significant covariates are the same.

Table 3. Summary Table for AFT Models

Covariates	Weibull	Log-Logistic	Log-Normal
Age	0.049*	0.042	0.033
Amputation	-0.171	-0.204	-0.177
Angioinv	-0.193	-0.238	-0.227
Chemo	0.247	0.267	0.274
diam (mm)	-0.015	-0.012	-0.013
esr1	0.885*	0.817*	0.784*
Grade	-0.651*	-0.631*	-0.6*
Histtype	-0.27	-0.238	-0.232
Hormonal	-0.06	-0.121	-0.01
Lymphinfil	0.513*	0.396	0.369
Posnodes	-0.034	-0.03	-0.03
Intercept	3.174	3.182	3.538
Shape Parameter	0.245	0.456	0.102
*tested at 99% level of confidence			

For the other 3 AFT models, the acceleration/deceleration time is affected by the effects of the multiple covariates measured via a log-linear model [12]. This also explains why coefficients/weights that comes with a certain covariate has an opposite sign compared to the Cox Models. This is due to positive coefficients DECELERATE the event time since the time is divided by the covariate/factor, indicating that this covariate/factor increases the mean/median survival time. Conversely a negative coefficient will ACCELERATE the occurrence of the event time which then reduce the mean/median survival time (sci-kit, Para 10). Though that may be the case, it is evident that regardless of varying coefficient and p-values for

all 6 models, estrogen 1 receptor expression values and cancer grade are the consistent significant predictors of Breast Cancer Survival as far as the data is concerned.

Table 4. Summary Table for Accuracy Metrics

	Classical Cox	Cox-Lasso	Cox-Ridge
C-Index	0.7706	0.7708	0.7709*
Partial AIC	752.3197	753.9405	752.6703*
	Weibull AFT	Log-Logistic	Log-Normal
C-Index	0.772	0.778	0.780*
AIC	618.634	612.981	608.822*

Across the 3 Cox variants, the Classical Cox regression shows the relatively least C-index indicating that the difference on the AIC of Classical Cox and Ridge-Cox is very minimal, which suggests that both models have good fit. However, Concordance Index [13] is the most reliable metric to be used as a model selection method.

In such events, right censoring (Occurrence of the event) must be considered rather than the model fit is considered. Higher C-index is the ideal metric since this is about the ratio between the occurrence of event and its non-occurrence (Concordance Index, para. 1), therefore leading to the decision of selecting the Cox-Ridge Regression as the most suitable model [14] for NKI Breast Cancer Survival Data despite its Partial AIC is 752.6703 [15].

For the 3 variants of the AFT models, contrasting results were acquired. Weibull AFT has the lowest C-index but the highest AIC while the Log-Normal AFT is the total inversion of the former's metric. This suggest that the latter model is to be used as the Parametric Survival Analysis model to be used as far as the C-index and AIC is concerned.

Table 5. Average Survival Times

Source	Average Survival Times	Confidence Intervals (99%)
From Raw Data	3.822 years	N/A
From Bootstrap (resampling = 2000 iterations)	4.476 years	[4.093 - 5.4088]

The table above shows the calculated Average Survival Times as well as their corresponding Confidence Intervals (for the bootstrap). It should be noted that there is a 0.654 year of differences between the 2 calculated averages. As far as the analysis shows, it can be inferred, that at a 99% level of confidence, that there is an estimated 4.476 years of estimated Survival Time after a certain patient has been diagnosed with Breast Cancer regardless of the patient profile without assuming the underlying distribution of the survival time.

The Hazard Function for the Cox-Ridge Regression is given by:

$$h(t) = h(t_0) * \exp(-0.063x_1 - 0.2862x_2 + 0.0109x_3 + 0.1974x_4 + 0.4192x_5 + 0.0204x_6 + 0.0444x_7 + 0.8454x_8 + 0.2476x_9 - 0.6579x_{10} - 1.1085x_{11})$$

Table 6. Predicted Survival (Upper and Lower 5 Subjects)

Subject No.	Time (in Years) - 18.34
228	0.961797355
118	0.958877968
7	0.958396389
144	0.956685445
66	0.954169466
15	0.007471557

155	0.006616058
30	0.003993136
95	0.002684222
131	1.19087E-13

In line with the estimated Survival Rate given by the Cox-Ridge model, this also suggests that those cases who share the same profile as with Patient 228 is estimated to survive by 96.18% after 18.34 years. Conversely, those individuals who share the same patient profile as with Patient 155 is assumed to survive by 000000000001191% after 18.34 years as far as the results is showing, further suggesting that these cases has the smallest likelihood of survival as far as the data is showing.

The Hazard Function for the Log-Normal AFT is given by:

$h(t)$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(t) - (0.033x_1 - 0.177x_2 - 0.227x_3 + 0.274x_4 - 0.013x_5 + 0.784x_6 - 0.6x_7 - 0.232x_8 + 0.369x_9 - 0.6579x_{10} - 0.03x_{11} + 3.538))^2}{2(0.102)^2}\right)$$

$$1 - \frac{\Phi(\log(t) - (0.033x_1 - 0.177x_2 - 0.227x_3 + 0.274x_4 - 0.013x_5 + 0.784x_6 - 0.6x_7 - 0.232x_8 + 0.369x_9 - 0.6579x_{10} - 0.03x_{11} + 3.538))}{0.102}$$

Table 7. Predicted Survival (Upper and Lower 5 Subjects)

Subject No.	Time (in Years)-18.34
144	0.937450397
7	0.936242412
228	0.931442656
118	0.919678237
134	0.918438023
95	0.071501812
100	0.064814634
30	0.064133071
155	0.061912506
131	0.010536416

In line with the estimated Survival Rate given by the Cox-Ridge model, this also suggests that those cases who share the same profile as with Patient 144 is estimated to survive by 93.75% after 18.34 years. Conversely, those individuals who share the same patient profile as with Patient 131 is assumed to survive by 1.1% after 18.34 years as far as the results is showing, further suggesting that these cases has the smallest likelihood of survival as far as the data is showing. It is very evident that the estimated Survival for each subject given by the 2 models is different, although similarity in the results is very evident.

3. CONCLUSION

Survival Analysis best shows how statistical analysis can literally save lives. A substantial amount of understanding on when to use a specific type of Survival Analysis model (whether univariate or multivariate as well as whether the assumptions for parametric, non-parametric, or semiparametric models are satisfied) is evident. As for what has been learned from this study, there is no single model that can act as a “one size fits all” test for predicting Breast Cancer Survival although it can be inferred, as far as this study is concerned, that Cox-Ridge Regression, an ideal basis for pure semi-parametric Machine Learning model, which in a way similar (based on Concordance Index) to what found out in their study. Though that may be the case, also said that slightest variation on the linear combination of covariates and changes in the partitioning of the training and testing dataset can greatly affect the metrics, C-index included. AFT models on the other hand are ideal if one wishes to apply the derivation of time ratio rather than the hazard ratio. In consideration of the metrics and coefficients acquired from the data, Log-normal

AFT outperforms its 2 variants giving us the idea that Log-normal, along with other AFT models are also “suitable for clinical research” Aalen (2000), which includes Cancer Survival Prediction as well. Due to the constraints encountered during the course of the study, individuals who might have an interest on this same topic or even healthcare practitioners in the field of Cancer prognosis and treatment can implement the same setup from this study to a different type of Cancer Survival (Lung, Colon, Skin, and Leukemia) to further validate and/or improve the current findings or future researchers can also localize the study by using clinical data with the same covariates used in this study from Filipina Breast Cancer patients to further exude relevance within the Breast Cancer situation in the country since Breast Cancer mortality in the Philippines is ranked #42 with a rate of 21.58 (as of this writing), according to worldlifeexpectancy.com.philippines-breast-cancer.

The said reason earlier is to also replicate and verify if the results will still hold true even if performed and tested on a different locale. Future researchers may also want to increase the number of observations to improve the performance of the Survival Prediction models to increase the C-index and further reduce the AIC.

Acknowledgment

The authors wish to profess unending and insurmountable gratitude and thanks for all of the guidance, help and support received from family, friends, loved ones and colleagues in the academe who in one way or another have contributed to making this study possible. Even though your names were not mentioned, your contributions regardless of the magnitude are truly appreciated and honored.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
John Edmon Alejandro Ganas	✓	✓	✓	✓		✓		✓	✓	✓	✓			
Peter John Berces Aranas		✓	✓		✓	✓		✓	✓		✓	✓	✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

Ethical Approval

Not Applicable.

Data Availability

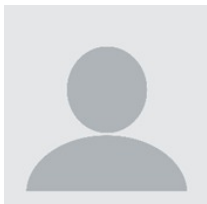
The data that support the findings of this study are available from the corresponding author upon reasonable request.

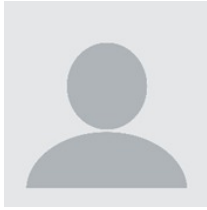
REFERENCES

- [1] 'The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser', The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics, vol. 21, no. 4, pp. 686-694], 1989.
- [2] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, 'Survival analysis part I: basic concepts and first analyses', Br. J. Cancer, vol. 89, no. 2, pp. 232-238, July 2003.
- [3] A. S. Allende, 'Concordance index as an evaluation metric', Medium, 12-Mar-2024. [Online]. Available: <https://medium.com/analytics-vidhya/concordance-index-72298c11eac7>.
- [4] V. Bewick, L. Cheek, and J. Ball, 'Statistics review 12: survival analysis', Crit. Care, vol. 8, no. 5, pp. 389-394, Oct. 2004. doi.org/10.1186/cc2955
- [5] M. Wang et al., 'Survival analysis of breast cancer subtypes in patients with spinal metastases', Spine (Phila. Pa. 1976), vol. 39, no. 19, pp. 1620-1627, Sept. 2014. doi.org/10.1097/BRS.0000000000000473
- [6] S. R. Cole, J. K. Edwards, A. Breskin, and M. G. Hudgens, 'Comparing parametric, nonparametric, and semiparametric estimators: The weibull trials', Am. J. Epidemiol., vol. 190, no. 8, pp. 1643-1651, Aug. 2021. doi.org/10.1093/aje/kwab024
- [7] E. Cai, 'On the C-statistics for Evaluating Overall Adequacy of Risk Procedures with Censored Survival Data Risk Prediction Procedures with Censored Survival Data', vol. 30, pp. 1105-1117, 2011. doi.org/10.1002/sim.4154
- [8] Chen, et. al (2020) A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction.
- [9] Cheng, et. al (2012) Prognosis of advanced hepatocellular carcinoma patients enrolled in clinical trial can be classified by current staging system 107: 1672-1677. doi.org/10.1038/bjc.2012.466
- [10] Ching, et.al (2022) Breast Cancer Survival Analysis Model 12(1971): 1, 10-12. doi.org/10.3390/app12041971
- [11] Z. Liang, 'Improved AIC Selection Strategy for Survival Analysis', Compute Stat Data Anal, vol. 52, no. 5, pp. 2538-2548, 2009. doi.org/10.1016/j.csda.2007.09.003
- [12] -. Li, 'Survival Analysis for a Breast Cancer Data Set', Advances in Breast Cancer Research, vol. 6, pp. 1-15, 2017.
- [13] R. Li, 'Machine Learning for Survival Analysis accelerated failure time models in high dimensions', vol. 41, pp. 933-949, 2018. doi.org/10.1002/sim.9264
- [14] S. Vishnubhata, 'Accelerated Failure Time Models', An Application in the Survival of Acute Liver Failure Patients in India, vol. 3, pp. 1-7, 2014.
- [15] S. Zhang, Application of Survival Analysis - Introduction and Discussion. ASQ Reliability Division, 2012.

How to Cite: John Edmon Alejandro Ganas, Peter John Berces Aranas. (2023). Evaluation of different survival analysis models for NKI breast cancer data. Journal Healthcare Treatment Development (JHTD), 3(2), 86-94. <https://doi.org/10.55529/jhtd.36.1.9>

BIOGRAPHIES OF AUTHORS

	John Edmon Alejandro Ganas, is a graduate student at the Polytechnic University of the Philippines-Manila, Philippines, pursuing advanced studies through its Graduate School. His research interests center on biostatistical modeling and survival analysis, with a focus on applying machine learning and parametric statistical methods to clinical and oncological datasets. His recent work investigates the comparative performance of semi-parametric and parametric survival models for breast cancer prognosis, contributing to the
---	---

	development of data-driven clinical prediction tools within the Philippine academic research landscape. Email: jeaganas@pup.edu.ph
	Peter John Berces Aranas , is a faculty member at the School of Statistics, University of the Philippines-Diliman, Philippines. His academic and research expertise lies in statistical theory and applied statistics, with particular emphasis on survival analysis, biostatistics, and predictive modeling. He has contributed to advancing statistical methodologies within the Philippine higher education system, mentoring graduate researchers in the rigorous application of both classical and contemporary statistical models to real-world clinical and epidemiological data. Email: pjbaranas@pup.edu.ph