Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

Evaluation of Different Survival Analysis Models for Nki Breast Cancer Data

John Edmon Alejandro Ganas[1*], Peter John Berces Aranas[2]

[1]Graduate School, Polytechnic University of the Philippines-Manila, Philippines.*
[2]School of Statistics, University of the Philippines-Diliman Philippines.

Email: [2]pjbaranas@pup.edu.ph
Corresponding Email: [1]jeaganas@pup.edu.ph*

Received: 28 May 2023 **Accepted:** 17 August 2023 **Published:** 01 October 2023

***Abstract**: 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. Highest Concordance Index and Lowest Akaike Information Criterion (AIC) are used as the basis of selecting the ideal Survival Analysis model as a template for the construction of the Survival Prediction model for NKI Breast Cancer Data. 6 Survival Analysis Models were used in this study. For the semi-parametric survival models, Classical Cox, Cox-Lasso, and Cox-Ridge Regressions. For the parametric models, 3 Accelerated Failure Time (AFT) models were implemented. These are: Weibull AFT, Log-logistic AFT, and Log-Normal AFT models. Right-censoring was performed in the data since it has been assumed that there are subjects which were not called back anymore for the entire, 18-year clinical trial where the data was taken from. A proportional hazards test was then performed to find out if the covariates in the data are fit to be modeled using Cox Regression and its derivatives. A test for the distribution on the time of event was also done to find whether it follows a specific distribution or not. This was done to verify the usability of the parametric survival analysis models on the data.*

It has been found out that in terms of Concordance Index and AIC, the Cox-Ridge Regression model outperforms its 2 other semi-parametric counterparts, having the least AIC of 752.6703 and Highest Concordance Index of 0.7709. As for the other 3 parametric models, Log-Normal AFT outperformed the Weibull AFT and Log-Logistic AFT models by a Concordance Index of 0.780 with a corresponding AIC of 608.822. This result also suggests that the time of event of the subjects is best fitted by Log-Normal Distribution. By comparing the 2, best-performing models, it has been reported that Log-Normal AFT outperforms Cox-Ridge Regressions, therefore suggesting to use this Parametric Survival Analysis Model as the basis for a Survival Prediction model suited for NKI Breast Cancer data.

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

Keywords: Survival Analysis Models, Concordance Index, Akaike Information Criterion (AIC), Breast Cancer, Semiparametric Models, Parametric Models.

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 Cadwell., et. Al (2017), "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 Gudi., et. al (2021), "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 Gudi., et. Al (2021), "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" (Flynn, 2012). 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 (Chen, et. al, 2020) 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 (Alagappan, et al. 2013).

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 Chen, et. al, (2020) 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.

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

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 (Simon, 2019) 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) (Barman and Saikia, 2017). 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 (Barman and Saikia, 2017). In other words, if the covariates exhibit a deceleration/acceleration on the response variable Survival Time, no strict assumptions are needed (Cox & Oakes, 1984). 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. (Chen, et. al, 2020).
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/FINDINGS

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			

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

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

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

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 (Archarya, 2012). 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 (Chen, et.al, 2020) 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 (Chen, et.al, 2020) for NKI Breast Cancer Survival Data despite its Partial AIC is 752.6703.

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

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.

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

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

$$h(t) = \frac{\frac{1}{\sqrt{2\pi\sigma t}}\exp\left(-\frac{(\log(t)-(0.033x_1-0.177x_2-0.227x_3+0.274x_4-0013x_5+0.784x_6-0.6x_7-0.232x_8+0.369x_9-0.6579x_{10}-0.03x_{11}+3.538)^2}{2(0.102)^2}\right)}{1-\frac{\Phi(\log(t)-(0.033x_1-0.177x_2-0.227x_3+0.274x_4-0013x_5+0.784x_6-0.6x_7-0.232x_8+0.369x_9-0.6579x_{10}-0.03x_{11}+3.538)}{0.102}}$$

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. CONCLUSIONS

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 Chen, et. al (2020) found out in their study. Though that may be the case, Chen, et. al (2020) and Xie (2014) 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

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

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 worldlifeecpectancy.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.

4. REFERENCES

1. Aban, et. al (2014) Survival Analysis and regression models, 21(4): 686-694J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
2. Altman, et. Al (2003) Survival Analysis Part 1: Basic concepts and first analyses, 89(2):232-238. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394262/M.
3. Allende, Alonso Silva (2019) Concordance Index as an Evaluation Metric. Available: https://medium.com/analytics-vidhya/concordance-index-72298c11eac7
4. Ball, et. al (2004) Statistics review 12: Survival Analysis, 8(5):289-394. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065034/
5. Bunger, et. al (2014) Survival Analysis of breast cancer subtypes in patients with spinal metastases, 39(19):1620-7. Available: https://pubmed.ncbi.nlm.nih.gov/24979144/
6. Breskin, et. al. (2021) Comparing Parametric, Nonparametric, and Semiparametric Models Semiparametric Estimators: The Weibull Trials, 190(8): 1643-1651. Available: https://pubmed.ncbi.nlm.nih.gov/33569578/
7. Cai, et. Al (2011) On the C-statistics for Evaluating Overall Adequacy of Risk Procedures with Censored Survival Data Risk Prediction Procedures with Censored Survival Data, 30(10):1105-1117.
8. Charan, Ravi (2020) The Cox Proportional Hazards Model: A Regression Model for Survival Data Regression Model for Survival Data. Available: https://towardsdatascience.com/the-cox-proportional-hazards-model-35e60e554d8f

Journal Healthcare Treatment Development
ISSN: 2799-1148
Vol: 03, No. 06, Oct-Nov 2023
http://journal.hmjournals.com/index.php/JHTD
DOI: https://doi.org/10.55529/jhtd.36.1.9

9. Chen, et. al (2020) A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction.
10. Cheng, et. al (2012) Prognosis of advanced hepatocelluar carcinoma patients enrolled in clinical trial can be classified by current staging system 107: 1672-1677.
11. Ching, et.al (2022) Breast Cancer Survival Analysis Model 12(1971): 1, 10-12.
12. Liang and Zou (2009) Improved AIC Selection Strategy for Survival Analysis. Compute Stat Data Anal; 52(5):2538-2548.
13. -Li, Hong (2017) Survival Analysis for a Breast Cancer Data Set. Advances in Breast Cancer Research, 6:1-15.
14. Li and Reddy (2018) Machine Learning for Survival Analysis accelerated failure time models in high dimensions, 41(6):933 - 949.
15. Vishnubhata, Sreenivas (2014) Accelerated Failure Time Models: An Application in the Survival of Acute Liver Failure Patients in India 3(6): 1-7.
16. Zajic, Alexandre (2019) Introduction to AIC - Akaike Information Criterion: Model selection without a validation or test set. Available: https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced
17. Zhang, Shaoang (2012) Application of Survival Analysis - Introduction and Discussion, ASQ Reliability Division