**Comparing Classical and Machine Learning Approaches to Predicting Breast Cancer Survival: A Log-Normal AFT Model vs. Neural Network Using the METABRIC Dataset**

David Wilczynski

College of Public Health, University of South Florida

**Author Note**

This paper has not been previously published or presented, and therefore, there are no conflicts of interest to report.

Correspondence concerning this article should be addressed to David Wilczynski, College of Public Health, University of South Florida, 13201 Bruce B. Downs Blvd, MDC 56, Tampa, FL 33612. Contact: [drwilczynski@usf.edu](mailto:drwilczynski@usf.edu)

**Comparing Classical and Machine Learning Approaches to Predicting Breast Cancer Survival: A Log-Normal AFT Model vs. Neural Network Using the METABRIC Dataset**

Breast cancer is quickly becoming one of the most common forms of cancer as it is estimated that 1 in 8 women will be diagnosed with breast cancer during her lifetime in the United States (American Cancer Society, 2024). While new methods of screening and genetic testing have become readily available within the last decade, it remains unclear why there has been such a spike in breast cancer as nearly half of cases diagnosed occur in women due to no specific risk factors other than sex and age (World Health Organization [WHO], 2024). The numbers have also shown that breast cancer rates are also increasing in men accounting for an estimated 1% of diagnoses and grim outcomes (American Cancer Society, 2024). In 2024, the World Health Organization estimated that 42,250 women and 530 men were expected to die in the US as a result of breast cancer or complications due to breast cancer (WHO, 2024).

Consequently, there has been an increased push for genetic and screening research to try and limit the increasing number of deaths from breast cancer through early diagnosis. A major breakthrough for targeted therapies occurred in 2009 when Parker et. al. published a paper which classified breast cancer tumors by their molecular subtype creating the PAM50 gene signature classification system. PAM50 [Prediction Analysis of Microarray 50] utilizes a tissue sample in order to measure the expression levels of 50 genes which were previously noted to play an important role in the development and growth of breast cancer (Parker et. al., 2009). A majority of the tested gene expression levels play a role in determining which receptors are expressed within the breast cancer tumor itself. The PAM50 molecular subtypes are broken down according to which receptors are expressed or not and the proliferation rate determined through the level of growth factor expressed. Parker et. al. described 6 molecular subtypes: Luminal A, Luminal B, HER2-Enriched, Basal-like, Normal-like, and Claudin-low. The PAM50 classification system went hand in hand with new targeted therapies and the development of treatment plans as it let the physician know if certain treatments like hormone therapies were a viable option given which receptors are most likely being expressed based on the PAM50 classification.

**Research Question**

With what seems to be a growing threat of breast cancer, there has been a push for increased methods of screening and testing to catch breast cancer earlier to reduce the chance of a grim outcome. Now that the PAM50 molecular subtyping classification system has been described, and there are now several effective treatments readily available for targeted approaches, the question arises: How does PAM50 molecular subtyping affect disease-specific survival time in breast cancer patients given key demographic and categorical covariates? Modeling the probability of disease-specific survival times based on PAM50 molecular subtyping could have an exponential clinical importance as it would allow for the creation of tumor specific treatments and screening plans based on the aggressiveness and chance of reoccurrence based on the PAM50 tumor classification. Including variables like age, tumor size and location, and possible treatments already tried could greatly increase the effectiveness of the model by accounting for possible bias and confounding.

With the research question now in place, a second methodological question arises: What is the most effective modeling framework to capture these complex relationships: classical statistical models or a more modern machine learning approach? Traditional statistical methods such as the Kaplan-Meier Estimator or Cox’s Proportional Hazard are the standard approach and offer interpretable and well-established techniques for modeling survival and time-to-event data. These models rely on statistical programs like R or SAS which utilize pre-specified distributions and linear relationships/regressions between the outcome and time-to-event variables while including space for potential covariates. On the other hand within the last decade, there has been an exponential growth in the availability and use of modern machine learning techniques such as a neural network which relies on splitting a dataset in two to first build and then train the model in order to detect nonlinear interactions and relationships that are usually overlooked by traditional statistics. Therefore, the objective of this analysis relies on building two models to serve as the representatives for traditional statistical modeling and modern machine learning techniques in order to answer the primary research question about PAM50 molecular subtyping and disease-specific survival time. The models will then be directly compared utilizing several model performance measures in order to test both the interpretability, fit, and clinical relevance within the fields of public health and medical research.

**Data Source and Variables Utilized**

For this analysis, a publicly available data set from the Molecular Taxonomy of Breast Cancer International Consortium [METABRIC] Study conducted in 2012 through the British Columbia Cancer Foundation and Canadian Breast Cancer Foundation BC/Yukon in partnership with Cancer Research UK was utilized as it contains the outcome and survival time of 1904 woman diagnosed with breast cancer from various clinics across Canada and the UK. This dataset was revolutionary for the field of oncological genetics as it contains the PAM50 molecular subtype and gene expression levels and patterns listed as nucleotide sequences for every tumor within the confines of the study. In terms of this survival analysis, survival time was measured in months (variable ⇒ overall\_survival\_months) and contains both censored and uncensored data. The outcome variable for the analysis was measured as disease-specific death (variable ⇒ death\_from\_cancer) coded as 1 = death from breast cancer or direct complication of breast cancer, and 0 = alive, censored, or death from other unrelated cause. The models were also given the task of incorporating a number of covariates known to have an impact on survival time. They are as follows: molecular subtype (pam50), age (age\_at\_diagnosis), tumor size (tumor\_size), tumor laterality (primary\_tumor\_laterality), chemotherapy (chemotherapy), hormone therapy (hormone\_therapy), radio therapy (radio\_therapy), breast surgery (type\_of\_breast\_surgery), and HER2 status (her2\_status). This allowed for the analysis of survival time based on PAM50 molecular subtype while incorporating how tumor size, location, and possible treatments utilized affected the overall outcome of survival.

**Methodology**

The first step within the survival analysis was determining which coding software to use for each model. For the traditional statistics model, R was used utilizing their survival analysis packages like survfit. For the modern neural network approach, Python was used utilizing a variety of Python libraries including Pandas and Pytorch. For the preliminary data exploring and analysis, a Kaplan-Meier [KM] Survival Curve was fit using both programs.

**Figure 1**

*A graph of a number of patients

AI-generated content may be incorrect.Survival Curves of PAM50 Molecular Subtypes in R*

**Figure 2**

*Survival Curves of PAM50 Molecular Subtypes in Python*

A graph of different colored lines

AI-generated content may be incorrect.

The next portion of the preliminary data analysis and preprocessing focused on checking the Cox Proportional Hazard assumptions against the data to see if a proportional hazard model could be used. By looking at the KM survival curves, it first became clear that the PH assumptions were violated by the data as the curves for each PAM50 molecular subtype crossed rather than running parallel. The PH assumptions were then further examined by running the cox.zph() procedure in R which resulted in the p-values for the PAM50 and full global models of <2e-16 and 3.8e-15 respectively. While steps were taken to try and mitigate this issue like including interaction terms and time-dependent variables within the PH models, the resulting p-values were consistently far too small to meet the PH assumptions, ruling out a Cox model for both coding programs.

**Traditional Statistics Model**

For the model representing traditional statistics coded in R, it was clear that a parametric model would be the best fit for this data based on the PH procedures mentioned above. Therefore, a Weibull Accelerated Failure Time (AFT) model was first fit to the data using first just the survival time, disease-specific outcome, and PAM50 molecular subtype. A second Weibull AFT model was then fit utilizing all the covariates mentioned to see which model was a better fit. Ultimately, the full model including all of the covariates was deemed to be the best fit by a Likelihood Ratio Test. The next step was then to determine which distribution was the right fit for the model, and therefore, an exponential, log-logistic, log-normal, and generalized gamma AFT model were also explored. Each model was then compared based on their AIC (see appendix) which determined that a Log-Normal AFT model was the best fit for the data.

**Figure 3**

*Q-Q Plot for the log-normal AFT model*

A graph with a line drawn on it

AI-generated content may be incorrect.

**Modern Machine Learning Approach**

For the modern approach utilizing new age data techniques, a neural network was coded in Python utilizing specific coding libraries including Pandas, Pytorch, and Seaborn. The data was first imported into Python and fit into the appropriate data frame using Pandas. As defined by a neural network, the dataset was split in two to have a subset to build the model and the remaining subset to train the model. After the traditional statistic model found a log-normal AFT as the best fit, the same distribution was utilized for the neural network. A log-normal AFT neural network was coded utilizing the negative log-likelihood loss. In order to train the model, the training data subset was run repeatedly through the model. Each pass is denoted as an epoch, and a total of 300 epoch were run in order to determine the overall loss per epoch as a measure of the model’s performance. Loss per epoch should decrease as the model is trained as more of the dataset is being utilized and accurately predicted by the model.

**Figure 4**

*Loss per Epoch of the Neural Network*

A graph of training loss

AI-generated content may be incorrect.

While the loss per epoch showed promising results, the resulting C-index of the log-normal AFT neural network was 0.355 which left room for improvement. The success of the model was then further evaluated by hyperparameter tuning by altering the optimizer and architectures which determine the depth of the model. The best performing log-normal AFT neural network model resulted in a C-index of 0.612. In order to determine if this was the best fit, the model was then switch over to a Weibull AFT model which resulted in a C-index of 0.663. Therefore, a Weibull AFT neural network was utilized for analysis and representative of modern machine learning techniques.

**Evaluation Metrics**

Now that both models were established, it was important to have a way to test the models against each other in terms of fit and over performance, as well as incorporating interpretability and overall generalizability. In order to compare the fit of the models, a combination of AIC, C-index, and time-dependent Brier Score was measured for each model which can be found in Results section and coding in the appendix.

**Results**

For the traditional statistics model, a log-normal AFT model was fit containing survival time in months, disease-specific death, and all the covariates: PAM50, age at diagnosis, tumor size, tumor laterality, surgery type, chemotherapy, hormone therapy, radio therapy, and HER2 status. This model was fit and tested against other combinations of models that contained a different mix or covariates or a different distribution and was determined to be the best fit using the likelihood ratio test, AIC, and Q-Q plots. The overall outcome of interest was set as survival time in months utilizing disease-specific death from breast cancer as the censoring variable to only code yes (1 = death from breast cancer) if the patient’s death was a direct cause of breast cancer or complications due to breast cancer. Utilizing R and several of its statistical coding packages, the coefficients, ETRs, and 95% Confidence Intervals [CI] were calculated and can be seen below in Table 1. The ETR values illustrate how each covariate directly impacts a patients estimated survival time while also showing how each PAM50 molecular subtype alters both positively or negatively overall survival time. In order to test the fit and comparison to the neural network model, there needs to be a benchmark or standard to use to compare the two models that was accessible and calculatable in both statistical programs.

**Table 1**

*Log-Normal AFT Coefficients and ETR Estimates*

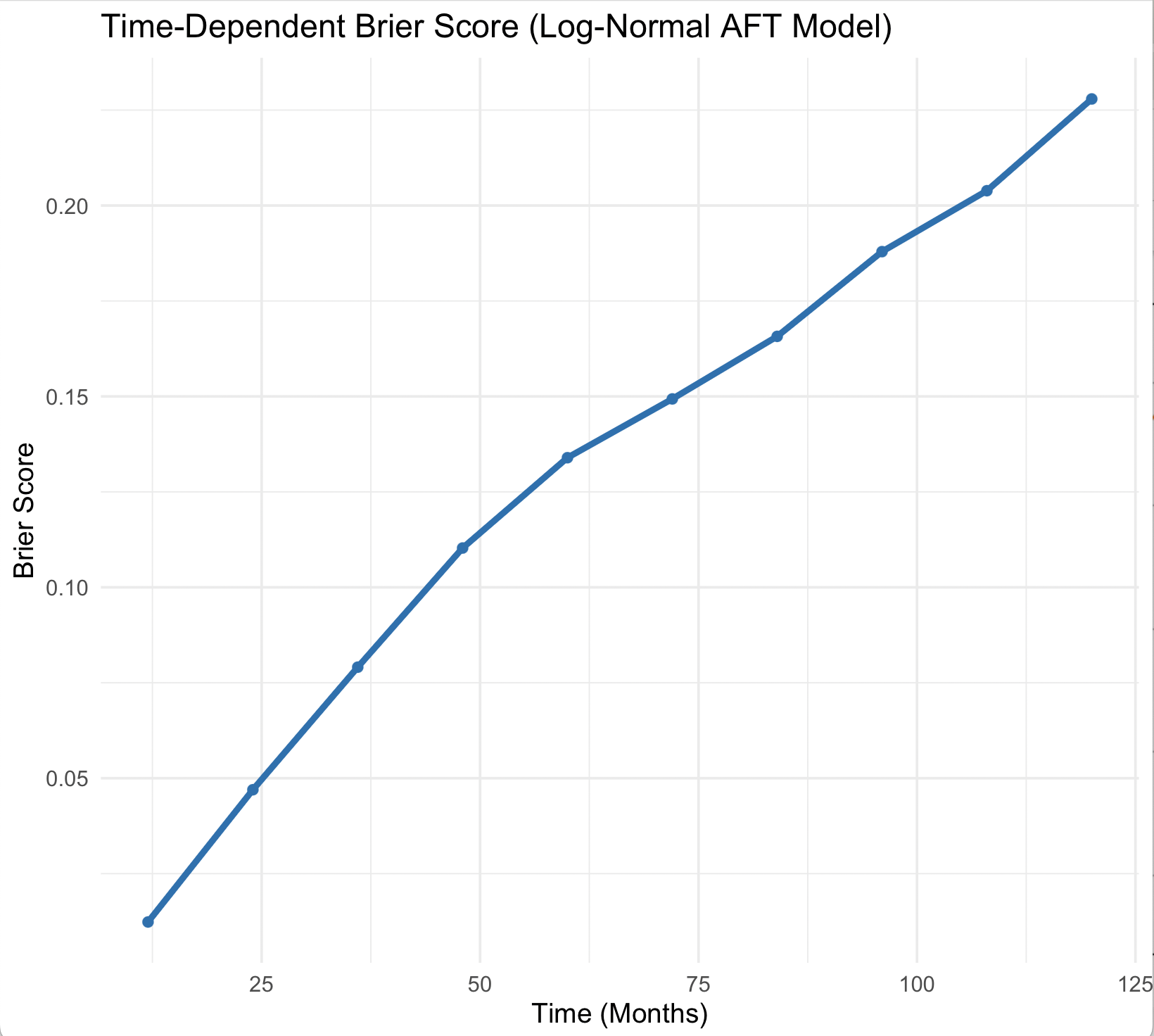
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Value | SE | p-value | | ETR | 95% CI |
| (Intercept) | 5.655 | 0.494 | < 2e-16 \* | 285.807 | | (108.584, 752.278) |
| PAM50 - Basal | -0.293 | 0.185 | 0.114 | 0.746 | | (0.519, 1.073) |
| PAM50 – Claudin-low | 0.152 | 0.190 | 0.422 | 1.164 | | (0.803, 1.688) |
| PAM50 – HER2 | -0.174 | 0.187 | 0.354 | 0.841 | | (0.582, 1.213) |
| PAM50 – Luminal A | 0.498 | 0.160 | 0.002 \* | 1.645 | | (1.202, 2.252) |
| PAM50 – Luminal B | -0.120 | 0.163 | 0.462 | 0.887 | | (0.644, 1.221) |
| Age at Diagnosis | -0.012 | 0.004 | 0.001 \* | 0.988 | | (0.981, 0.995) |
| Tumor Size | -0.013 | 0.003 | 2.0e-07 \* | 0.987 | | (0.982, 0.992) |
| Breast Conserving Surgery | 0.820 | 0.425 | 0.054 | 2.269 | | (0.988, 5.215) |
| Mastectomy (surgery) | 0.466 | 0.420 | 0.267 | 1.594 | | (0.699, 3.632) |
| Left Laterality | 0.543 | 0.172 | 0.002 \* | 1.721 | | (1.229, 2.410) |
| Right Laterality | 0.650 | 0.174 | 0.0001 \* | 1.916 | | (1.363, 2.694) |
| Chemotherapy | -0.618 | 0.115 | 7.2e-08 \* | 0.539 | | (0.430, 0.675) |
| Hormone Therapy | 0.025 | 0.091 | 0.781 | 1.026 | | (0.858, 1.226) |
| Radio Therapy | -0.078 | 0.099 | 0.430 | 0.925 | | (0.761, 1.123) |
| Positive HER2 Status | -0.390 | 0.128 | 0.002 \* | 0.677 | | (0.526, 0.870) |
| Log(scale) | 0.313 | 0.031 | < 2e-16 \* | 1.387 | | (1.285, 1.454) |

*Note.* Statistical significance is denoted by \*, time is measured in months, and the reference category: for PAM50 is Normal, for breast surgery is None, and for Laterality is Bilateral.

Therefore, it was determined that in order to compare the two models a combination of their AIC, C-index, and time-dependent Brier Scores would be used to test the overall fit and collective picture of the data in order to determine if traditional statistics or modern machine learning techniques had a larger impact and clinical relevance. For the log-normal AFT model, the model reported an AIC of 8291.572 and a C-index of 0.696. The time-dependent Brier Score plot can be seen below in Figure 5. When looking at the Brier plot, a good fit of a model is shown by limiting or having a low Brier Score throughout the entire model. The log-normal AFT model demonstrated a good fit within the terms of the Brier Score by peaking at 0.25 at 125 months.

**Figure 5**

*Time-Dependent Brier Score for Log-Normal AFT Model in R*



**Neural Network in Python**

The modern machine learning approach consisted of coding a neural network in Python utilizing several Python libraries designed for data manipulation and model architecture. The first model as previously mentioned was designed as a log-normal AFT neural network as that distribution showed promising results in the R output for traditional statistics. However, a low C-index of 0.355 highlighted that the model might not be the best fit. Therefore, the first possible solution was to alter the model’s hyperparameters including the optimizer and architecture. This helped increase the fit of the model resulting in a best fit C-index of 0.612. Altering the model to fit a Weibull distribution on the other hand resulted in a C-index of 0.663 which resulted in being the best fit for the model as only parametric methods could be utilized. The coefficients, ETR values, 95% CI for the ETRs, and p-vales can be seen below in Table 2.

**Figure 6**

*Time-Dependent Brier Score Weibull AFT Neural Network in Python*

*A graph with a line

AI-generated content may be incorrect.*

In terms of the performance metrics, as previously stated the Weibull AFT neural network resulted in a C-index of 0.663. A final AIC value was calculated in Python resulting in an AIC of 6303.415. The time-dependent Brier Score Plot as seen in Figure 6 did not illustrate promising results as the Brier Score peaked at 50 months with a value of 50.

**Table 2**

*Weibull AFT Neural Network Coefficients and ETR Estimates*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Value | SE | p-value | | ETR | 95% CI |
| (Intercept) | 7.627 | 0.303 | < 0.0000 \* | 2053.39 | | (1134.47, 3716.62) |
| PAM50 – HER2 | -0.193 | 0.171 | 0.257 | 0.824 | | (0.589, 1.152) |
| PAM50 – Lum A | 0.309 | 0.159 | 0.0529 | 1.361 | | (0.996, 1.861) |
| PAM50 – Lum B | -0.152 | 0.159 | 0.338 | 0.859 | | (0.630, 1.172) |
| PAM50 – Normal | -0.163 | 0.189 | 0.388 | 0.849 | | (0.587, 1.230) |
| PAM50 – Claudin-low | 0.285 | 0.187 | 0.128 | 1.329 | | (0.921, 1.918) |
| Age at Diagnosis | -0.018 | 0.004 | 1.22e-05 \* | 0.982 | | (0.975, 0.990) |
| Tumor Size | -0.010 | 0.002 | 6.04e-07 \* | 0.990 | | (0.986, 0.994) |
| Type of Breast Surgery | -0.351 | 0.106 | 9.56e-04 \* | 0.704 | | (0.571, 0.867) |
| Tumor Laterality | 0.088 | 0.083 | 0.290 | 1.092 | | (0.928, 1.285) |
| Chemotherapy | -0.619 | 0.124 | 5.42e-07 \* | 0.538 | | (0.423, 0.686) |
| Hormone Therapy | 0.017 | 0.094 | 0.852 | 1.017 | | (0.847, 1.223) |
| Radio Therapy | -0.195 | 0.002 | 6.14e-02 | 0.823 | | (0.671, 1.009) |
| HER2 Status | -0.365 | 0.131 | 5.15e-03 \* | 0.694 | | (0.537, 0.896) |
| Rho | 0.122 | 0.040 | 2.38e-03 \* | 1.130 | | (1.044, 1.223) |

*Note.* Statistical significance is denoted by \* and time is measured in months.

**Interpretations: Log-Normal AFT Model**

Several of the coefficients within the log-normal AFT model showed statistical significance. The PAM50 molecular subtype Luminal A was the first coefficient to show statistical significance with a p-value of 0.002 and an ETR value of 1.645 indicating that the molecular subtype Luminal A is associated with a 65% longer estimated survival time compared to the reference group which was set as molecular subtype Normal-like. Age at diagnosis also displayed statistical significance with a p-value of 0.001 and an ETR of 0.988 indicated that with every 1 unit increase in age (+ 1 year) the estimated survival time decreases by 1.2%. An extremely similar result was found with tumor size which had a significant p-value of 2.0e-07 and an ETR of 0.987 indicating that for every one unit increase in tumor size the estimated survival time decreases by 1.3%. Although displaying borderline significance with a p-value of 0.054, breast conserving surgery resulted in an ETR of 2.269 meaning more than double the survival time compared to the reference group which was no surgery. Both left and right laterality showed statistical significance with p-values of 0.002 and 0.0001 respectively. Left laterality had an ETR of 1.721 indicating a 72.1% increase in survival time, while right laterality had an ETR of 1.916 indicating almost a 92% longer survival time compared to the reference group which was bilateral tumors. Of the other treatment types included in the model, chemotherapy was the only one to show statistical significance with an extremely low p-value of 7.2e-08 and an ETR of 0.539 meaning a 46.1% shorter survival time for those who received chemotherapy. Finally, a positive HER2 status had a p-value of 0.002 and an ETR of 0.677 indicating that a positive HER2 is associated with a 32% shorter estimated survival time.

While not statistically significant, both hormone therapy and radio therapy resulted in an ETR value of almost 1 (1.026 and 0.925) indicating no real change in survival time. The molecular subtype Basal displayed an ETR of 0.746 which was the lowest of all the PAM50 types indicating basal expression levels resulted in the shortest estimated survival time of all the PAM50 subtypes. Ultimately, age, tumor size, chemotherapy, and HER2 positive status significantly reduces survival time, while Luminal A subtype, breast-conserving surgery, and right or left tumor laterality increase survival time.

**Interpretations: Weibull AFT Neural Network and Best Fit Comparison**

Several of the coefficients within the Weibull AFT Neural Network also illustrated statistical significance. None of the PAM50 molecular subtypes indicated statistical significance, but Luminal A was borderline with a p-value of 0.0529 and an ETR of 1.361 which could possibly indicate a 36.1% longer survival. Age at diagnosis and Tumor size both resulted in ETRs extremely close to 1 at 0.982 and 0.990 respectively. This indicates that every increasing year reduces survival time by 1.8% and each unit increase of tumor size (mm) reduces survival by 1%. No breast surgery resulted in an ETR of 0.704 indicating a 29.6% shorter survival time compared to removing the tumor. Chemotherapy had an ETR value of 0.538 which is associated with a 46.2% shorter survival time, which could be an indicator of bias within the dataset as chemotherapy is a last resort treatment for breast cancer meaning sicker patients receive chemotherapy. Finally, HER2 status was the last coefficient to show statistical significance with an ETR value of 0.694 showing that positive HER2 status reduces survival time by 30.6%.

In terms of the better fit when comparing traditional statistics to a more modern machine learning approach, traditional statistics utilizing the log-normal AFT model had the better fit for the data. In terms of C-index, the log-normal AFT had a C-index of 0.696 compared to 0.663 for the Weibull AFT neural network coded in Python. This is further illustrated by looking at the time-dependent Brier Score plots as shown above in Figures 5 and 6. The Brier Score for the log-normal AFT model peaked at 125 months with a value of approximately 0.25. The Weibull AFT neural network did significantly worse peaking at a Brier Score of around 50 at 50 months. When looking back at the models that held the same distributions, the log-normal AFT model from R still held the better fit when comparing the C-index values as previously mentioned it had a C-index of 0.696, while the log-normal AFT Neural Network from Python had a best fit C-index of 0.612 after altering the hyperparameters. The AICs however show conflicting results as the Weibull AFT neural network had an AIC of 6303.415 compared to the log-normal AFT model in R which had an AIC of 8291.572 which associates the neural network as the better fit. However, this result is biased as a foundational step in building a neural network is to split the dataset into the building and training blocks meaning that only a portion of the dataset was used to build the model while the other portion was used to train the model. This means the fit of the model was only tested against the portion of the data used to build the model.

**Discussion**

The two models illustrated the current capabilities of traditional statistical coding and modern machine learning approaches in their respective coding programs. The traditional statistics model represented as a log-normal AFT model demonstrated superior fit and performance as measured by the performance indicators particularly the C-index and time-dependent Brier Score. While the neural network achieved the lower AIC, this is not a fair metric to use for comparison as previously stated due to the nature of a neural network which only uses a subset of the overall dataset in order to build the model potentially inflating or overfitting the model during the training phase. With the current Python libraries available for machine learning approaches, it is necessary to have an extremely large dataset to build and train the model in order to have enough data for both procedures while also still being able to alter the hyperparameters to achieve the best fit for the model. In terms of interpretability, the log-normal AFT model offered greater interpretability especially for the PAM50 molecular subtypes which was the main research question of this analysis. The traditional statistics model gave clear and interpretable ETRs which illustrated direct impacts of estimated survival time. On the other hand, the Python model while designed to look for and examine more non-linear and complex relationships struggled to interpret the given METABRIC dataset illustrating a hindrance to the overall clinical relevance. A neural network in its given iteration and coding capacities needs a large dataset in order to catch the nuances in the data. Given Public Health and medical research, this greatly impacts the usability of neural networks as the availability and collection of data within these fields is simply not enough. This makes sense given the fact that currently neural networks for analysis are mainly only used with the field of tech and finance where there are large dataset readily available and other AI programs and processes that aid in the creation and analysis of neural networks. Another benefit to traditional statistical modeling is its reproducibility which is also a known issue for neural network. Traditional statistical coding relies on the given distribution or statistical formulas to build the model, and as an added benefit seed numbers are often used and required when necessary. In Python however, the utilized Python libraries like Pytorch used for neural network building make it hard to maintain reproducibility as the model training portion of the coding can greatly alter the results and epoch loss as it changes every time the program is re-run. However, this is not to say that this will always be an issue or rule out machine learning techniques from public health and medical research altogether as this is a relatively new field with improvements and exponential advancements being made consistently. A team in China is actually working on making survival analysis more accessible in machine learning and neural network approaches by building a new Python library that can handle the statistical assumptions and requirements necessary for survival analysis. Given the current iterations, a traditional statistic approach and modeling still hold more clinical relevance and interpretability when it comes to survival analysis and medical research where decisions must be evidence based and reproducible.

This dataset also came into the modeling with its own issues. This dataset is quite biased when it comes to the nature of the patients and severity of breast cancer. METABRIC was a study designed to build a genetic database of gene expression levels and molecular sequencing, and therefore, more severe cases of breast cancer were cataloged, like those with large or bilateral tumors. The sample is also entirely female, which is not representative of the distribution of breast cancer in the population. The patients within the METABRIC sample also represented an older demographic with a mean age of 61.1 and range of (22, 96). If this analysis were to be run again, a larger more comprehensive dataset would be needed in order to allow both traditional statistical modeling and a more modern machine learning approach a chance to accurately depict and model the data. Ideally, the best fit for the data and models would be the same distribution across both models as well.

**Conclusion**

In conclusion, the purpose of this analysis was to test how PAM50 molecular subtypes of breast cancer while including potential covariates affect estimated disease-specific survival time while also examining whether traditional statistical modeling or a modern machine learning approach provided a better fit for the data keeping in mind overall interpretability and impact on clinical relevance. As previously stated, the log-normal AFT model coded in R and representing traditional statistical modeling proved to be a better fit when comparing C-index and time-dependent Brier Scores. The log-normal AFT model also offered greater interpretability of the overall model with the provided output and calculated ETRs, p-values, and 95% CI. This allowed for direct examination of how each covariate and the model as a whole affect the estimated disease-specific survival time of breast cancer patients. In summary, age, tumor size, chemotherapy, and HER2 positive status significantly reduces survival time, while Luminal A subtype, breast-conserving surgery, and right or left tumor laterality increase survival time compared to the reference group for each covariate. On the other hand, the Weibull AFT Neural Network initially showed promising results with the calculated coefficients and plots representative of the model. However, the model fell short when comparing the performance metrics specifically C-index and time-dependent Brier Score. It was also illuminated that the neural network had lower interpretability and clinical relevance as the model relies on a process that needs a much larger dataset in order to detect the subtle nuances and both linear and non-linear relationships. Overall, the traditional statistical methods remain the gold standard when it comes to public health and medical research as they maintain high interpretability and reproducibility which are essential in making evidence based and educated informed decisions. While this model focused on a neural network, there are other machine learning approaches that also might be able to fit a survival analysis in their given iterations like deep learning. It is also important to highlight that the best fit for each model resulted in different parametric distributions. Advancements, however, are constantly being made within the machine learning field with the development and integration of AI and programming languages. While the current state of machine learning is not up to par for clinical analysis, there could be future advancements or hybrid models which allow for a more accurate representation and direct model for analysis. As it currently stands, traditional statistics is far more suited to handle survival analysis with interpretable while reproducible results for medical research given the data.

**References**

American Cancer Society. (2024). Breast Cancer Facts and Figures. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2024/breast-cancer-facts-and-figures-2024.pdf>

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group, Langerød, A., Green, A., Provenzano, E., Aparicio, S. (2012). The Genomic and Transcriptomic Architecture of 2,000 Breast Tumors Reveals Novel Subgroups. *Nature*, *486*(7403), 346–352. <https://doi.org/10.1038/nature10983>

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., & Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, *27*(8), 1160–1167. <https://doi.org/10.1200/JCO.2008.18.1370>

World Health Organization. (2024). Breast Cancer. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>

**Appendix**

**Full Python Model and Coding**: <https://github.com/david-wilczynski/Portfolio-DW/blob/103d916ecb4ba9a3e2cab88cf86d81fd3b6d3640/SurvivalData-NeuralNetwork/NN_Survival.ipynb>

**Figure 7 and 8**

*Initial fitting of Cox PH Model for both Univariate (PAM50) and Multivariate*

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 9**

*Validity Test for Cox PH*

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 10 and 11**

*Initial Weibull AFT models both univariate (PAM50) and multivariate*

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

**Figure 12**

*Likelihood Ratio testing and AIC that led to selection of Log-Normal AFT model*

A computer code with blue text

AI-generated content may be incorrect.

A computer code with blue text

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

A close-up of numbers

AI-generated content may be incorrect.

**Figure 13**

*Log-Normal AFT Model Output* A screenshot of a computer program

AI-generated content may be incorrect.

**Figure 14**

*ETR values for Log-Normal*

A computer screen shot of a computer code

AI-generated content may be incorrect.

**Figure 15**

*Calculation of respective ETR 95% CIs*

A screenshot of a computer code

AI-generated content may be incorrect.