Title: Clinical Factors Predicting Progression-Free Survival in Triple Negative Breast Cancer

Introduction

Triple negative breast cancer (TNBC) is defined as a type of breast cancer absent the following common receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Pathologically, it can be confirmed through immunohistochemistry indicating ER of less than 1%, PR of less than 1%, and HER2 receptor score of 0 to 1+. TNBC is typically seen as a more aggressive cancer type that is harder to treat and has a higher likelihood of recurrence. Because TNBC is considered a high-risk phenotype the threshold for offering adjuvant chemotherapy is lower compared to non-TNBC patients.[1]

Adjuvant or neoadjuvant chemotherapy refers to the use of cytotoxic chemotherapy after having breast cancer surgery. The benefits of this therapy include a decreased risk of recurrence and mortality. The risks of chemotherapy are quite severe, including but not limited to nausea, vomiting, hair loss, immunosuppression that can lead to a higher risk of infections, and early cognitive impairment.[2]

TNBC accounts for 12-17% of all breast cancers and occurs more frequently in Black and Hispanic women that are less than 50 years of age than in similarly aged women of other racial or ethnic groups.[3] TNBCs are also characterized by a higher expression of Ki-67, a nuclear antigen that is typically used as a marker of active cell proliferation, and a higher frequency of BRCA1 mutations, a gene associated with an increased risk of breast cancer and is a common mutation seen in breast cancer patients. As stated before, the likelihood of recurrence or mortality is higher in TNBC than in other breast cancer types, with a high risk of this outcome within 5 years of diagnosis.[4,5] Given the heterogeneous nature of TNBC, the differing rates of TNBC in different demographic groups, and the limited treatment options for this disease as well as the cost and chemotoxicity associated with chemotherapy; further research into the clinicopathological features and sociodemographic factors associated with prognosis and patterns of recurrence is necessary in order to increase survival outcomes.

Previous studies have supported that adjuvant therapy after primary breast surgery is associated with higher survival. Using a larger subset of the data we are utilizing, Newman et al. have previously concluded that node-negative TNBC associated with T1c tumors had positive

outcomes when following the National Comprehensive Cancer Network guidelines for the recommendation of adjuvant therapy. However, cases of T1a and T1b did not have significant improvement when receiving chemotherapy versus not.[6,7] While there is previous research on what the thresholds for the recommendation of chemotherapy should be and that chemotherapy is associated with higher survival in TNBC patients post breast surgery, there is minimal detailed and consistent research on what demographic and clinical factors are associated with survival after receiving this treatment.

The tumor, node, and metastasis (TNM) stage has consistently been established as a prognostic factor for breast cancer but its predictive value is less clear in TNBC.[8] The ratio of positive lymph nodes to the total number of removed lymph nodes has been previously speculated to be another potential prognostic factor but few clinical studies have thoroughly examined its predictive value.[9] The utilization of antigen Ki-67 as a prognostic factor, however, has been supported in several clinical studies.[4] In our analyses, we aim to explore the rate of progression-free survival (PFS) in TNBC patients and fill in the gap of knowledge in prognostic factors previously indicated by building a Cox regression model and using a random forest classifier.

Methods

The data analyzed is from a retrospective study conducted by Weill Cornell Medicine. Patient records were selected from patients treated in Manhattan at Weill Cornell Medicine. Patients had to meet the following inclusion criteria: seen between December 1999 and June 2018, age 18 and older, pathologically confirmed TNBC, and if HER2/neu immunohistochemistry is 2+ then they must be negative for amplification by fluorescence in situ hybridization (FISH). The patients also had to have tumors in the pathologic stage T1N0 and have undergone surgical therapy without receiving neoadjuvant treatment. Additionally, clinical records had to be complete, especially verified hormone receptor and/or HER2 status and confirmation of not having adjuvant chemotherapy.

The data was organized for further analysis by ensuring the assignment of the correct variable type (factor for categorical variables, numeric for continuous variables, and date-time for date variables), transforming variables, and detecting missing values. The following variables were re-categorized by combining smaller-sized groups into larger groups to aid in data analysis: race/ethnicity, spanish origin, the surgical procedure of primary site, regional LN surgery (LN biopsy/removal versus other), histology (idc/ilc versus other), pT stage of primary surgery, pN stage of primary surgery, post-op/adjuvant XRT status (adjuvant XRT received versus not), the reason for no radiation (recommended versus not recommended), T class - clinical, N class - clinical, clinical stage group, T class - pathologic, N class - pathologic, pathologic stage group, ki67, ER expression, and PR expression.

Asian and Pacific Islander race/ethnicity groups (encoded as 5-12) were collapsed into one group while Hispanic/Latinx and Indigenous populations (4, 13) were re-categorized as Other (15). Spanish origin was also re-categorized as Spanish (1, 3, 4, 9), Non-Spanish (10), and Unknown (2). For the N class variables, they were relabeled as node-positive (1) and node-negative (0) and for T class, the pathologic was kept as stage 1-4 while clinical was re-grouped to T1/T2 and T3/T4. The clinical stage group and pathologic stage group were simplified to stage 1-4 and 2-3, respectively. Similarly, the primary pT stage was transformed to the levels T1, T2, and T3/4 and the primary pN stage to N0, N1, and N2/3. For the surgical procedure of the primary site, total mastectomies were combined (40-49) and due to few patients in each group, all other procedures other than total mastectomy and lumpectomy/excisional biopsy were labeled as 'other'. Lastly, the percentage expression of ki67 (G1 - less than or equal to 5%, G2 - greater than 5% and less than or equal to 20%, and G3 - greater than 20%), ER (less than 1%, greater than 1%), PR (less than 1%, greater than 1%) were stratified into categories.

In data selection, the analysis inclusion criteria for predictors were the following: missingness less than 70% (imputation with larger missing values may result in bias), meaningful (non-duplicated variables, cannot be calculated from other variables, not appropriate in PFS e.g. local recurrence, distance recurrence) and non-date variables (single date point gives no information). After the removal of predictors that did not meet the previously stated criteria, single imputation methods（mode, mean) were applied to missing values. Observations with recurrence/death date prior to diagnosis date were also removed (patient ID 622, 479, and 809). After these steps, the data was split randomly into training (70% of patients) and test sets (30% of patients). Three seeds were utilized to test for sensitivity (25, 35, and 45).

Analysis Methods

Survival (KM/Cox PH)

For survival analysis, we applied the Kaplan–Meier (KM) method and the Cox Proportional-Hazards (PH) model to analyze 'time-to-event' data. The outcome in KM analysis was the recurrence of breast cancer or mortality, and the time to it was called survival time. The Cox PH was used to simultaneously evaluate the effect of several predictors on survival , specifically how certain factors influence the rate of recurrence/mortality at a particular point in time. This rate is commonly referred to as the hazard rate.

In order to determine if the survival distributions were significantly different between distinct levels of an independent variable, log-rank tests were conducted on all variables that met the inclusion criteria. The predictors that had hypothesis tests resulting in a p-value less than 0.2 were considered significant predictors, with the null hypothesis as there is no difference between the survival curves of the various levels of a predictor and the alternate hypothesis being there is a difference.

One assumption of the Cox PH model is that the relative hazards remain constant over time with different levels of the predictor. For categorical variables, this assumption was visually checked by verifying if the Kaplan-Meier progression-free survival curves crossed at any time point. For continuous variables, this assumption was examined by visually inspecting the functional form by plotting the continuous variable against martingale residuals of the null Cox proportional hazards model using the ggcoxfunctional() function from the survminer library and then checking the linearity of the fitted line. Plots of the quadratic form of the continuous variable and the square root of the continuous variable were generated and inspected in the same manner in order to evaluate if the functional form was non-linear. Chi-squared tests were also conducted for continuous variables with the null hypothesis that the proportional hazards (PH) assumption is appropriate and the results were deemed significant if the p-value was less than .05. All variables that did not meet the proportional hazards assumption were not considered potential predictors.

For the variable selection of our final Cox PH model, both backward stepwise selection using the stepAIC() function from the MASS library and elastic net regression using the glmnet library were utilized. The models that were selected from these methods were compared using a concordance (c-index), a measure of the proportion of pairs of patients whose observed and predicted outcomes match among all possible pairs in which one patient experiences the outcome of interest (recurrence/death) and the other one does not, and area under the curve (AUC), which is the probability that a randomly selected pair of patients (one truly positive and one truly negative for the outcome) will be correctly ordered by the model. This AUC measure was visualized with a receiver operating characteristic (ROC) curve, which plots the true positive rate versus the false-positive rate at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both two rates. The AUC measure comes from the entire two-dimensional area underneath the entire ROC curve from the coordinates (0,0) to (1,1). A larger AUC value indicates better predictive performance, with .5 implies no discrimination between classes while .1 indicates perfect discrimination.

Backward stepwise selection begins with a model that contains all potential variables of interest (variables without high missingness, with clinical significance, and meeting the PH assumption) and then removes the least significant variable one by one until a stopping criterion or the intercept-only model is reached. The stopping criteria were determined by an Akaike information criterion (AIC) value which measures how well a particular model fits the data.

The elastic net regression fits a regularized regression to a generalized linear model using the penalty term: $(1 - \alpha)||\beta||_2^2 + \alpha||\beta||_1$, where alpha balances the l_1 penalty of lasso regression and the l_2 penalty of ridge regression. This regression method also uses a tuning parameter λ that determines the amount of shrinkage of the coefficient terms induced by the penalty function. First we fixed α as .5, the midpoint value of α which can range from 0 to 1, and chose the best λ

from a grid of 100 values from .01 to 1e10 via cross-validation (10 folds) and a deviance loss function. We then did a separate search with parallelism for the best α and the associated best λ value using cross-validation (20 folds) and a deviance loss function. A larger number of folds were used to minimize the variability of the resulting α and λ. From these three potential models (backwards selection, elastic net regression with α = .5 and a search for λ, and the elastic net with a search for α and λ), the model with the highest concordance and AUC on the training and test data set was chosen as the final model.

Random Survival Forest

The random survival forest (RSF) model, an implementation of Breiman's random forests (RF) method, was applied on right-censored survival data. Compared to other survival approaches, this method has the advantage of not requiring restrictive assumptions, such as proportional hazards, and it also handles non-linearity and interaction identification difficulties automatically.[10] We applied RSF-related functions contained the grf R package.[11] First, variable selection is determined based on the importance of predictors, which is the weighted sum of the frequency of one variable split at each depth in RSF. Then according to Nicolò et. al., the tuning parameters of the number of variables sampled at every split and the minimum number of observations in a terminal node were identified by maximizing the c-index computed from out-of-bag (OOB) data.[12] We calculated the OOB concordance with the mortality score detailed in Ishwaran et al.[10] For consistency, c-index on the test set is used to measure the predictive power of RSF.

Results

The outcome variable was PFS, and survival time was the date at diagnosis to the last follow-up/death/recurrence in days. Three incorrect observations were removed: one was left-censored (patient ID 479), one had zero survival time (patient ID 809), and the other had no diagnosis time (patient ID 622). After data cleaning and processing, we had 267 patients and 45 variables (Appendix, Table 1b).

In patients treated with a second breast conserving surgery with intraoperative radiotherapy, 3-year PFS was significantly associated with tumor size and stage: tumor size pathology (p-value = .006), clinical stage group (p-value = .006), T class - clinical (p-value = .033), clinical T stage (p-value = .028), clinical N stage (p-value = .013). Additionally the predictors neoadjuvant chemotherapy status (p-value < .001), CPM (p-value = .025), first treatment modality (p-value < .001), nodal metastases diagnosis (p-value = .008), age at diagnosis (p-value - .049), and genetic testing (< .001) were significant (Appendix, Table 2).

Survival (KM/Cox PH)

First, the univariate K-M curve for our variable of interest chemotherapy was plotted and the results suggest that the survival rate for patients with and without recurrence/mortality is significantly different (Figure 1, p-value < .0001). Then, univariate survival curves and log-rank tests were implemented on the 45 predictors that met the inclusion criteria, and 24 predictors that did not violate the PH assumption and were statistically significant (p-value < 0.2) were selected. The categorical variables selected were as follows: MRI screen, Index Tumor Status, past contralateral, T class - clinical, clinical stage group, T class - pathologic, pathologicc stage group, reason for no radiation, radiation/surgery sequence, high grade disease status, HER2/neu (FISH), HER2 status, first treatment modality, SLN biopsy status, ALND, clinical T stage, clinical N stage, pT stage, neoadjuvant chemotherapy , postop/adjuvant CTX, and postop/adjuvant XRT. The linearity of the functional form of three numeric predictors was verified: age at diagnosis, tumor size pathology, regional lymph node examination. These 24 features (21 categorical variables and 3 continuous variables) were included in the multivariable Cox model as predictor variables, recurrence or morality as the event, and survival time as the time.

Next, we applied two methods for variable selection: backward selection and elastic net regression. For backward selection, 20 variables were outputted as predictor variables, but there evidence of overfitting, indicated by the difference in c-index of the training and test sets (Table 3). For elastic net selection, 6 variables were selected as predictor variables with tuning parameters $\alpha = 0.95$ and $\lambda = 0.0588$ (Appendix, Table 4 and Figure 2). Both two models satisfied the PH assumption of the Cox PH model. The model of backward selection has a c-index of 0.640 and AUC of 0.621 on the test set; the model of elastic net selection has a c-index of 0.695 and AUC of 0.604 (Appendix, Figure 2). Thus, we prefer the model from elastic net selection as the final model, which includes the predictors reason for no radiation, high grade disease status, first treatment modality, clinical stage group, T class - pathologic, and clinical N stage. The univariate KM plots of these 6 predictors are included below (Figure 2).

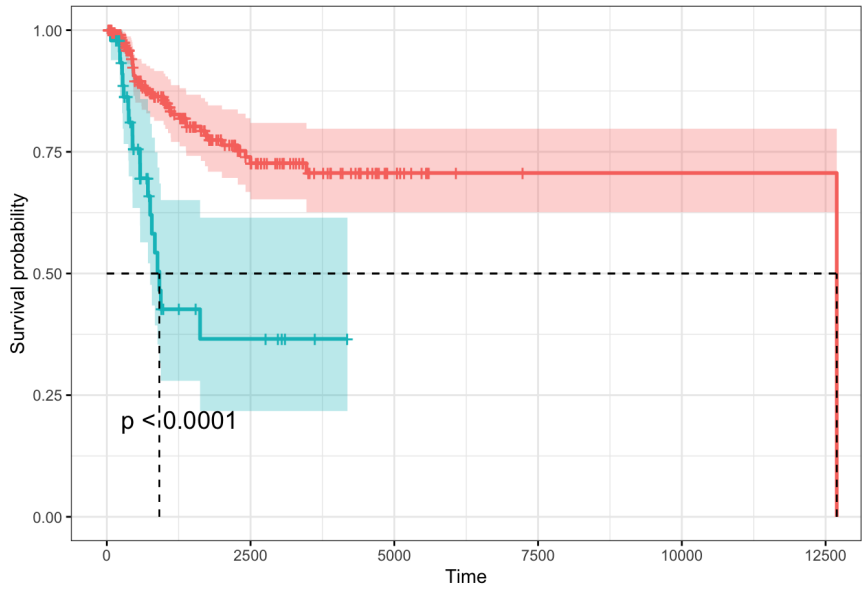
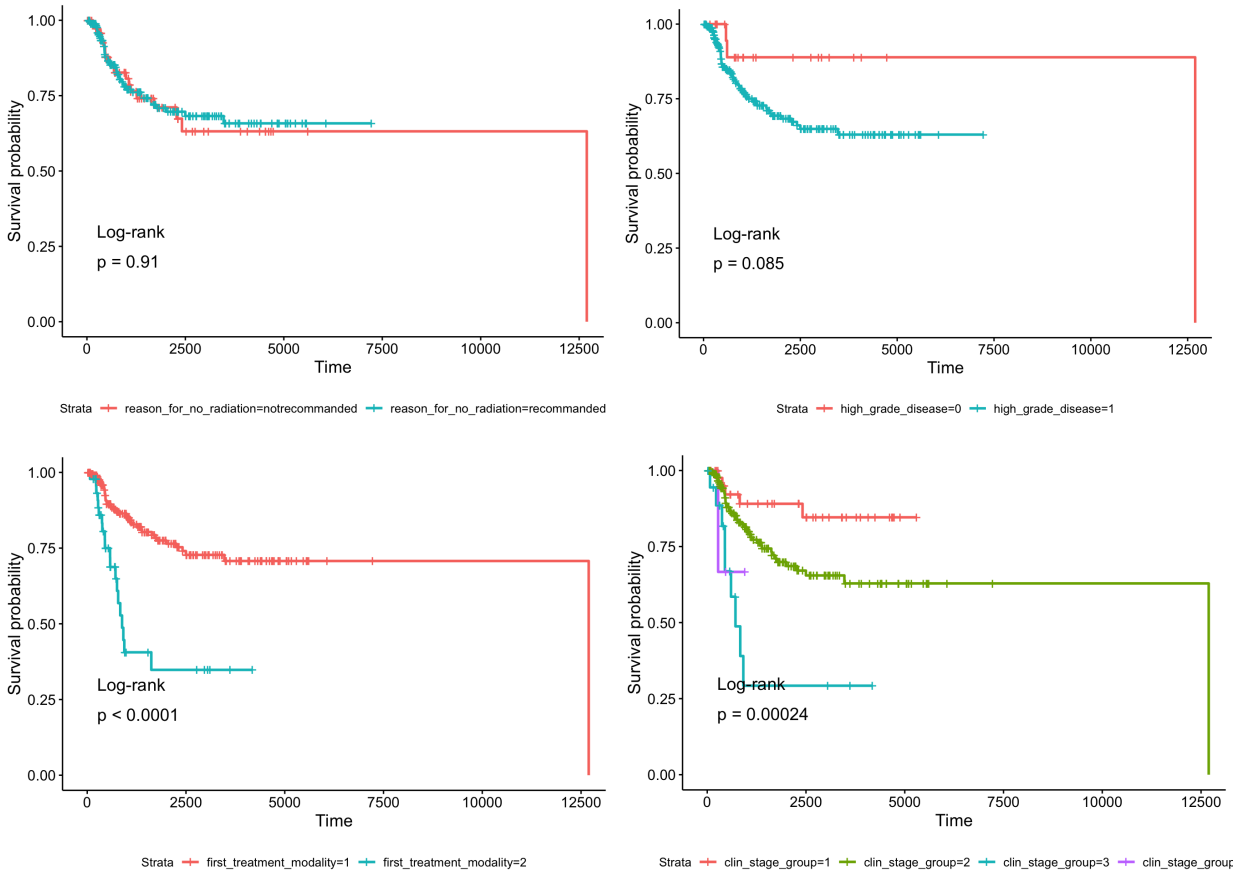


Figure 1: K-M Survival Curve for Chemotherapy.



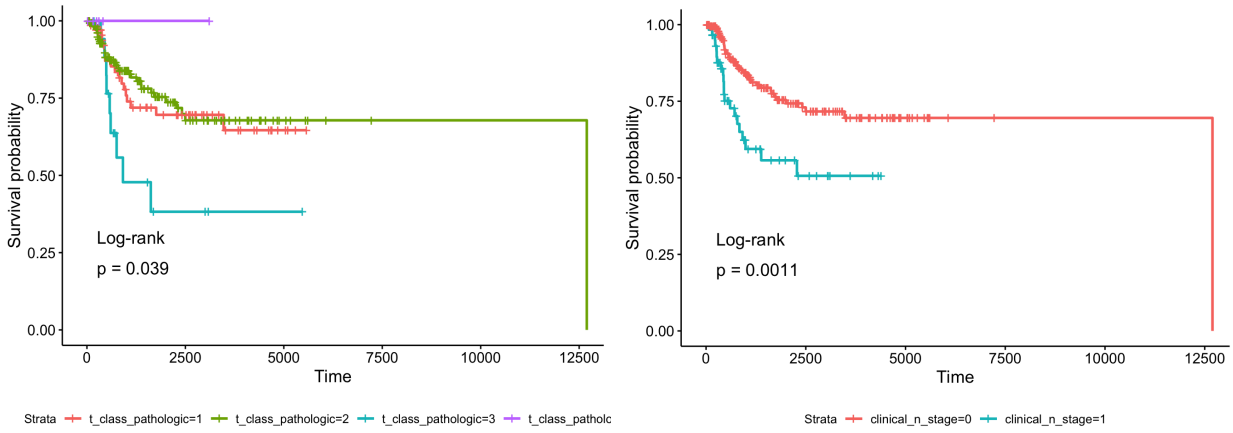


Figure 2: Univariate K-M Survival Curves of Elastic Net-Selected Variables. Predictors: for reason for no radiation, high grade disease status, first treatment modality, clinical stage group, T class - pathologic, and clinical N stage.

Random Survival Forest

Besides regression, one machine learning technique was applied to the progression-free survival patient data: RSF. In the RSF, we recognized the variables with high predictive power along with the accuracy of the prediction. At first, a design matrix with dummy variables was created from predictors that are factors, and the dimension of the predictor space increased from 43 to 61. As random forest can be a feature selection method and measure feature importance, we performed the RSF on the training set and obtained results of the exclusion of the variable MRI screen-detected. All the other variables proceeded to parameter tuning.

We tuned two parameters, the number of variables at each split and the minimum number of data points in a leaf. Both optimal parameters are picked to maximize OOB concordance. Figure 3 shows the process of this optimization, where the model decides to have 17 variables sampled at each split and at least 11 observations in each tree leaf. In addition, the top ten predictors from the optimal model are specified. In this case, age at diagnosis has the highest predictive power followed by neoadjuvant chemotherapy and clinical N stage. In total 37 variables contributed to predicting recurrence/mortality. For the optimal model, the c-index on the test set is 0.654 (95% CI, 0.52 to 0.79). For the predicted survival curve for each observation in Figure 4, it is noted that the recurrence or mortality probabilities at 3 years ranges from 0.18 to 0.3.

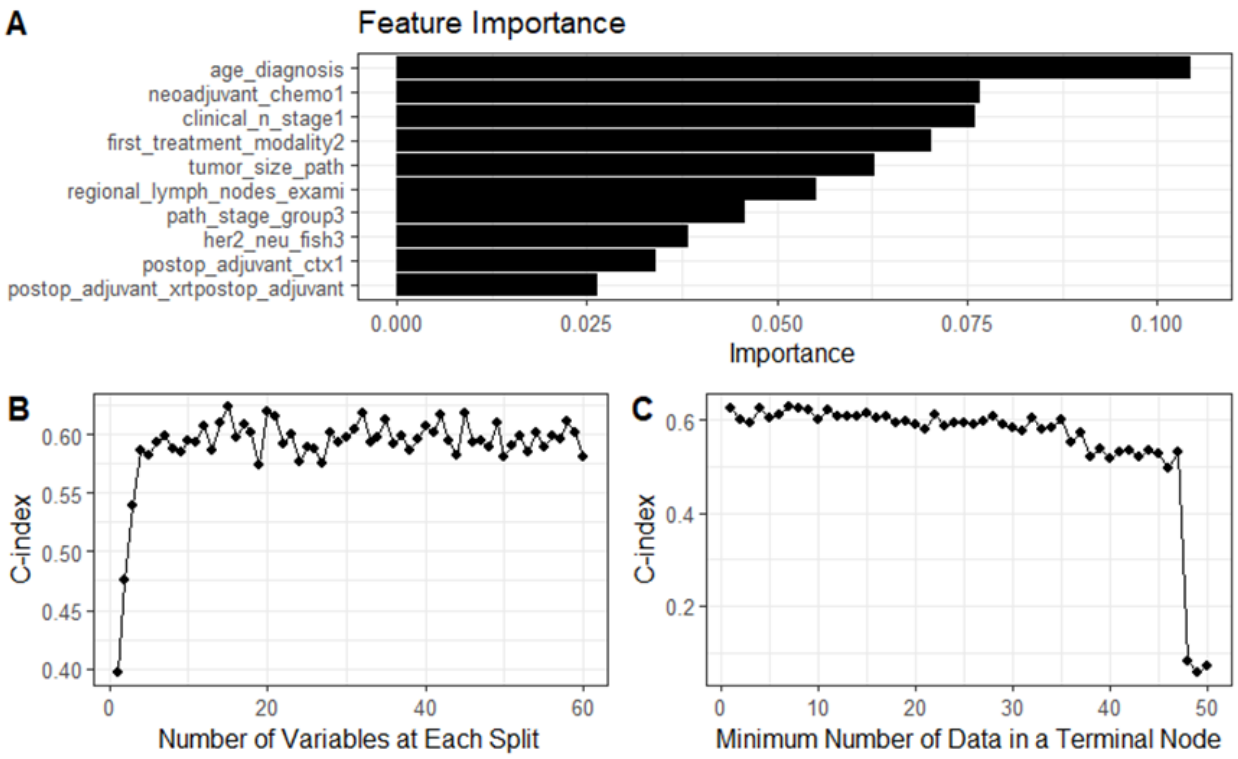


Figure 3: Survival Random Forest and Feature Selection. (A) Variable Importance of Top 10 Predictors. (B) Harrell's C-index under different numbers of variables at each split. (C) Harrell's C-index under the different minimum number of data points in a terminal node.

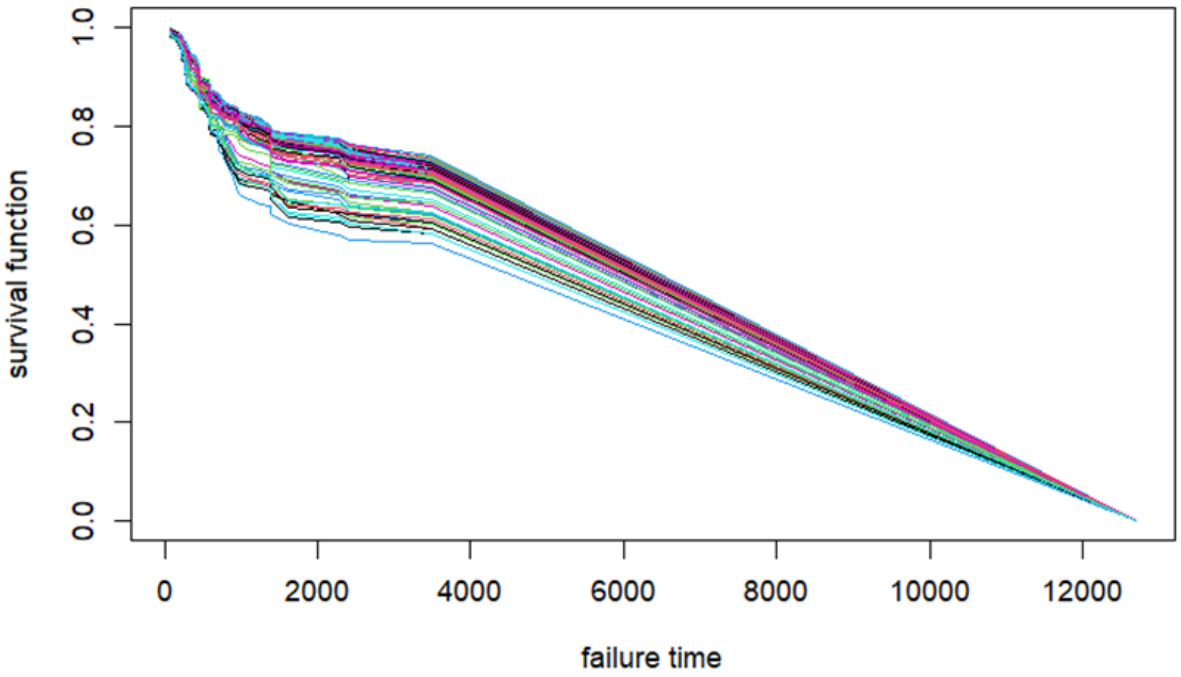


Figure 4: Predicted Survival Curve for all 77 observations in the test set.

Discussion

To identify potential factors that are associated with progression-free survival of stage 2-3 TNBC patients, we constructed two predictive models: an elastic net-selected Cox PH regression and a random survival forest. Their predictive powers were measured by the c-index of the 77 patients of the test set, and the c-index of the training set was also considered in a later comparison between methods. As for the Cox PH model, we tried two common variable selection approaches, the backward selection selected 20 among 24 predictors and resulted in a concordance of 0.64 (95% CI, 0.504 to 0.776). In contrast, tuning the α and λ values in the elastic net provide fewer numbers of predictors (6) and have slightly higher concordance (0.695, 95% CI, 0.577 to 0.814). Furthermore, the concordance in the training set using backward selection has a high value of 0.808, suggesting that the large number of selected variables can lead to overfitting. As a result, we decided to use the elastic net selected-variables for the Cox regression. In addition, the random survival forest gives a c-index of 0.654 (95% CI, 0.52 to 0.79), which is also lower than the previous method. Therefore, the elastic net Cox regression is the advised method for PFS prediction in this study.

The features that have predictive power according to elastic net Cox regression include clinical stage group, pathologic T class, reasons for no radiation, high-grade disease status, first treatment modality, and clinical N stage. Within these factors, the first treatment modality and clinical N stage are significant at a 5% level. This hints that node-positive patients have generally longer survival compared to node-negative patients. However, the RSF includes more factors in the prediction, and except for high-grade disease, all the other factors from the elastic net selection are contained. RSF also suggests that age at diagnosis, adjuvant chemotherapy status, and tumor pathology size are important in prediction as well.

Table 6. Comparison of Predictive Performance of Models

Method	C-Index	95% CI - Lower Bound	95% CI - Upper Bound
Backwards Selection	.640	.504	.776
Elastic Net	.695	.577	.814
RSF	.654	.520	.790

Moreover, we also performed the Chi-square test, Fisher's exact test, and Wilcoxon test to assess the association between 43 inputted predictors and 3-year PFS status (Appendix, Table 2). We found that the majority of significant variables are consistent with the two models. For example, age is on average 4 years older in the recurrence group and tumor size is on average 0.3 cm smaller in the progression-free group. Nevertheless, this does not take into account censored patients and several variables violate the proportional hazard assumption such as needle biopsy-proven nodal metastases at diagnosis. Fasano et. al. used multivariable analysis on the overall survival of TNBC patients and concluded that the benefit of adjuvant chemotherapy in node-negative TNBC along with other significant factors such as tumor size and grade 3 disease.[7] Thus, the PFS analysis in our study is efficacious because our outcome contains these factors.

Due to the random split of training and validation sets, we also tested the sensitivity of randomness on the results with two more seeds in R (35 and 45, original seed was 25). Although there are a few differences in selecting variables in each method, the results are similar to what we have described. For one of the other versions of the data split between training and test sets (i.e. using one of the seeds other than 25), the resulting c-index is close for RSF and elastic net Xox regression with RSF being slightly higher. However, in general, the elastic net cCx method has better performance in prediction.

There are several limitations would like to address in the patient data and our methodology. First, when handling the missing values, we used single imputations with mode and means for simplicity. In this case, the multivariate imputation methods might have been a better choice. Second, the categorization in data cleaning may not be appropriate. Many predictors had groups with few patients and this caused problems in calculating models and measures in our initial survival analysis. To avoid this issue, we decided to combine several factors. For instance, 15 race/ethnicity groups were shrunk to 5 groups (WA, AA, UK, Asian Pacific, and Other) and T class clinical was regrouped as the two categories T1/T2 and T3/T4. Therefore, there is information loss in some cases as well as difficulty in interpretting several of these predictors and their respective categories. Thirdly, collinearity between categorical variables is another important assessment in descriptive data analysis that can be involved in this study. We also identified three patients that did not satisfy the criterion of having a logical survival time (diagnosis before recurrence/mortality) which resulted in their removal in our analysis. Ideally, we would verify the information of these patients again with the electronic health record. Finally, other machine learning methods are worth attempting such as XGBoost— a method that is more flexible, efficient, and can handle missing values with its built-in features.

Reference

1. Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P., & Narod, S. A. (2007). Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *13*(15 Pt 1), 4429–4434. https://doi.org/10.1158/1078-0432.CCR-06-3045

2. Haidinger, R., & Bauerfeind, I. (2019). Long-Term Side Effects of Adjuvant Therapy in Primary Breast Cancer Patients: Results of a Web-Based Survey. *Breast care (Basel, Switzerland)*, *14*(2), 111–116. https://doi.org/10.1159/000497233

3. Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A., & Caggiano, V. (2007). Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry. *Cancer*, *109*(9), 1721–1728. https://doi.org/10.1002/cncr.22618

4. Urru, S., Gallus, S., Bosetti, C., Moi, T., Medda, R., Sollai, E., Murgia, A., Sanges, F., Pira, G., Manca, A., Palmas, D., Floris, M., Asunis, A. M., Atzori, F., Carru, C., D'Incalci, M., Ghiani, M., Marras, V., Onnis, D., Santona, M. C., … Orrù, S. (2018). Clinical and pathological factors influencing survival in a large cohort of triple-negative breast cancer patients. *BMC cancer*, *18*(1), 56. https://doi.org/10.1186/s12885-017-3969-y

5. Reis-Filho, J. S., & Tutt, A. N. (2008). Triple negative tumours: a critical review. *Histopathology*, *52*(1), 108–118. https://doi.org/10.1111/j.1365-2559.2007.02889.x

6. Gradishar, W. J., Anderson, B. O., Abraham, J., Aft, R., Agnese, D., Allison, K. H., Blair, S. L., Burstein, H. J., Dang, C., Elias, A. D., Giordano, S. H., Goetz, M. P., Goldstein, L. J., Isakoff, S. J., Krishnamurthy, J., Lyons, J., Marcom, P. K., Matro, J., Mayer, I. A., Moran, M. S., … Kumar, R. (2020). Breast Cancer, Version 3.2020, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network : JNCCN*, *18*(4), 452–478. https://doi.org/10.6004/jnccn.2020.0016

7. Fasano, G. A., Bayard, S., Chen, Y., Varella, L., Cigler, T., Bensenhaver, J., Simmons, R., Swistel, A., Marti, J., Moore, A., Andreopoulou, E., Ng, J., Brandmaier, A., Formenti, S., Ali, H., Davis, M., & Newman, L. (2022). Benefit of adjuvant chemotherapy in node-negative T1a versus T1b and T1c triple-negative breast cancer. *Breast cancer research and treatment*, *192*(1), 163–173. https://doi.org/10.1007/s10549-021-06481-4

8. Park, Y. H., Lee, S. J., Cho, E. Y., Choi, Y., Lee, J. E., Nam, S. J., Yang, J. H., Shin, J. H., Ko, E. Y., Han, B. K., Ahn, J. S., & Im, Y. H. (2011). Clinical relevance of TNM staging system according to breast cancer subtypes. *Annals of oncology : official journal of the European Society for Medical Oncology*, *22*(7), 1554–1560. https://doi.org/10.1093/annonc/mdq617

9. Vinh-Hung, V., Verkooijen, H. M., Fioretta, G., Neyroud-Caspar, I., Rapiti, E., Vlastos, G., Deglise, C., Usel, M., Lutz, J. M., & Bouchardy, C. (2009). Lymph node ratio as an alternative to pN staging in node-positive breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *27*(7), 1062–1068. https://doi.org/10.1200/JCO.2008.18.6965

10. Ishwaran, H., Kogalur, U., Blackstone, E., & Lauer, M. (2008). Random Survival Forests. *The Annals of Applied Statistics, 2*(3), 841-860.

11. Julie Tibshirani, Susan Athey, Erik Sverdrup and Stefan Wager (2022). grf: Generalized Random Forests. R package version 2.1.0. https://CRAN.R-project.org/package=grf

12. Nicolò, C., Périer, C., Prague, M., Bellera, C., MacGrogan, G., Saut, O., & Benzekry, S. (2020). Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer. *JCO Clinical Cancer Informatics, 4*(4), 259-274.

Appendix

Table 1a. Raw TNBC patient data (see attachment)

Table 1b. Clean TNBC patient data

Characteristic	0, N = 206[1]	1, N = 61[1]
Race		
African American	27 (13%)	7 (11%)
Asian/Pacific Islander	18 (8.7%)	3 (4.9%)
Other	17 (8.3%)	4 (6.6%)
UK	18 (8.7%)	5 (8.2%)
White American	126 (61%)	42 (69%)
Spanish Origin		
Non-Spanish	151 (73%)	49 (80%)
Spanish	17 (8.3%)	3 (4.9%)
Unknown	38 (18%)	9 (15%)
Age at Diagnosis	56 (14)	53 (14)
Mammo Screen-Detected		
0	135 (72%)	37 (76%)
1	52 (28%)	12 (24%)
Missing	19	12

Mammo Occult		
0	186 (98%)	47 (96%)
1	4 (2.1%)	2 (4.1%)
Missing	16	12
MRI Screen-Detected		
0	182 (96%)	50 (100%)
1	7 (3.7%)	0 (0%)
Missing	17	11
Index Tumor Status		
1	186 (92%)	50 (88%)
2	16 (7.9%)	7 (12%)
Missing	4	4
Past Ipsilateral		
0	197 (98%)	55 (96%)
1	5 (2.5%)	2 (3.5%)
Missing	4	4
Pat Contralateral		

0	188 (93%)	51 (89%)
1	14 (6.9%)	6 (11%)
Missing	4	4
Tumor Size - By Pathology	2.77 (1.99)	3.80 (4.19)
missing	28	18
Surgical Procedure of Primary Site		
Lum_ex_biopsy	86 (42%)	26 (43%)
Other	52 (25%)	12 (20%)
Total Mastectomy	68 (33%)	23 (38%)
Regional LN Surgery		
Other	16 (7.8%)	6 (9.8%)
Regional LN Surgery	190 (92%)	55 (90%)
Regional Lymph Nodes Examined	10 (9)	12 (9)
Missing	54	16
T Class - Clinical		
T1/T2	140 (90%)	28 (78%)

T3/T4	16 (10%)	8 (22%)
missing	50	25
Clinical Stage Group		
1	43 (28%)	5 (14%)
2	100 (65%)	20 (57%)
3	10 (6.5%)	9 (26%)
4	2 (1.3%)	1 (2.9%)
Missing	51	26
T Class - Pathologic		
1	56 (27%)	18 (30%)
2	133 (65%)	33 (55%)
3	9 (4.4%)	9 (15%)
4	6 (2.9%)	0 (0%)
Missing	2	1
Pathologic Stage Group		
2	168 (82%)	42 (69%)
3	38 (18%)	19 (31%)

Hormone Therapy		
0	178 (91%)	54 (92%)
1	17 (8.7%)	5 (8.5%)
Missing	11	2
Reason For No Radiation		
Not recommended	60 (32%)	19 (32%)
Recommended	126 (68%)	40 (68%)
Missing	20	2
Radiation/Surgery Sequence		
1	69 (36%)	24 (41%)
2	124 (64%)	34 (59%)
Missing	13	3
Metastases Diagnosis		
0	156 (80%)	47 (82%)
1	39 (20%)	10 (18%)
Missing	11	4
Histology		

idc/ilc	191 (95%)	58 (98%)
Other	10 (5.0%)	1 (1.7%)
Missing	5	2
Grade of Differentiation (Pre-8th)		
2	26 (13%)	3 (5.1%)
3	163 (80%)	50 (85%)
9	14 (6.9%)	6 (10%)
Missing	3	2
High Grade Disease Status		
0	24 (12%)	3 (5.5%)
1	169 (88%)	52 (95%)
missing	13	6
LVI Status		
0	96 (57%)	18 (42%)
1	73 (43%)	25 (58%)
Missing	37	18
HER2/neu (P/N/UK)		

1	135 (67%)	42 (72%)
2	23 (11%)	5 (8.6%)
3	45 (22%)	11 (19%)
Missing	3	3
HER2/neu (FISH)		
1	111 (58%)	27 (48%)
2	11 (5.7%)	3 (5.4%)
3	31 (16%)	13 (23%)
4	39 (20%)	13 (23%)
Missing	14	5
HER2 Status		
1	150 (73%)	49 (80%)
2	56 (27%)	12 (20%)
First Treatment Modality		
1	177 (87%)	37 (65%)
2	26 (13%)	20 (35%)
Missing	3	4

Attempt at Lumpectomy Status		
0	88 (43%)	26 (45%)
1	117 (57%)	32 (55%)
Missing	1	3
Mastectomy Surgery		
0	101 (50%)	24 (41%)
1	103 (50%)	35 (59%)
Missing	2	2
CPM		
0	163 (80%)	44 (76%)
1	40 (20%)	14 (24%)
Missing	3	3
SLN Biopsy Status		
0	39 (19%)	16 (28%)
1	164 (81%)	41 (72%)
Missing	3	4
ALND		

0	97 (48%)	22 (39%)
1	107 (52%)	35 (61%)
Missing	2	4
Clinical T Stage		
1	54 (34%)	9 (23%)
2	88 (56%)	22 (56%)
3	9 (5.7%)	4 (10%)
4	6 (3.8%)	4 (10%)
Missing	49	22
Clinical N Stage		
0	116 (75%)	14 (40%)
1	39 (25%)	21 (60%)
Missing	51	26
Primary pT Stage		
T1	46 (26%)	9 (23%)
T2	120 (69%)	26 (67%)
T3/4	9 (5.1%)	4 (10%)

Missing	31	22
Primary pN Stage		
N0	81 (46%)	13 (33%)
N1	74 (42%)	18 (46%)
N2/3	21 (12%)	8 (21%)
Missing	30	22
Neoadjuvant Chemotherapy Status		
0	176 (87%)	38 (66%)
1	27 (13%)	20 (34%)
Missing	3	3
Postop/Adjuvant CTX		
0	24 (13%)	11 (20%)
1	158 (87%)	44 (80%)
Missing	24	6
Postop/Adjuvant XRT		
None	51 (30%)	20 (38%)
Postop/Adjuvant	120 (70%)	32 (62%)

Missing	35	9
Adjuvant Endocrine Therapy Status		
0	161 (88%)	51 (91%)
1	21 (12%)	5 (8.9%)
Missing	24	5
Genetic Testing		
0	89 (56%)	23 (49%)
1	70 (44%)	24 (51%)
Missing	47	14
Survival Time	1,776 (1,627)	1,015 (1,663)

[1]n (%); Mean (SD)

Table 2. Hypothesis testing of three-year PFS

Characteristic	0, N = 122[1]	1, N = 145[1]	p-value[2]
Race			0.2
African American	15 (12%)	19 (13%)	
Asian/Pacific Islander	12 (9.8%)	9 (6.2%)	
Other	7 (5.7%)	14 (9.7%)	

UK	6 (4.9%)	17 (12%)	
White American	82 (67%)	86 (59%)	
Spanish Origin			0.094
Non-Spanish	96 (79%)	104 (72%)	
Spanish	11 (9.0%)	9 (6.2%)	
Unknown	15 (12%)	32 (22%)	
Age at Diagnosis	53 (12)	57 (15)	0.049
Mammo Screen-Detected			0.4
0	90 (74%)	113 (78%)	
1	32 (26%)	32 (22%)	
Mammo Occult			0.2
0	121 (99%)	140 (97%)	
1	1 (0.8%)	5 (3.4%)	
MRI Screen-Detected			0.7
0	118 (97%)	142 (98%)	
1	4 (3.3%)	3 (2.1%)	
Index Tumor Status			0.12

1	115 (94%)	129 (89%)	
2	7 (5.7%)	16 (11%)	
Past Ipsilateral			>0.9
0	119 (98%)	141 (97%)	
1	3 (2.5%)	4 (2.8%)	
Pat Contralateral			0.053
0	117 (96%)	130 (90%)	
1	5 (4.1%)	15 (10%)	
Tumor Size - By Pathology	2.82 (2.53)	3.09 (2.20)	0.006
Surgical Procedure of Primary Site			0.2
Lumpectomy or Excisional Biopsy	45 (37%)	67 (46%)	
Other	35 (29%)	29 (20%)	
Total Mastectomy	42 (34%)	49 (34%)	
Regional LN Surgery			0.7
Other	11 (9.0%)	11 (7.6%)	
Regional LN Surgery	111 (91%)	134 (92%)	

Regional Lymph Nodes Examined	11 (8)	11 (7)	0.9
T Class - Clinical			0.033
T1/T2	116 (95%)	127 (88%)	
T3/T4	6 (4.9%)	18 (12%)	
Clinical Stage Group			0.006
1	27 (22%)	21 (14%)	
2	92 (75%)	105 (72%)	
3	3 (2.5%)	16 (11%)	
4	0 (0%)	3 (2.1%)	
T Class - Pathologic			0.3
1	38 (31%)	36 (25%)	
2	77 (63%)	92 (63%)	
3	6 (4.9%)	12 (8.3%)	
4	1 (0.8%)	5 (3.4%)	
Pathologic Stage Group			0.070
2	102 (84%)	108 (74%)	
3	20 (16%)	37 (26%)	

Hormone Therapy			>0.9
0	112 (92%)	133 (92%)	
1	10 (8.2%)	12 (8.3%)	
Reason for No Radiation			0.8
Not Recommended	37 (30%)	42 (29%)	
Recommended	85 (70%)	103 (71%)	
Radiation/Surgery Sequence			0.4
1	39 (32%)	54 (37%)	
2	83 (68%)	91 (63%)	
Metastases Diagnosis			0.008
0	108 (89%)	110 (76%)	
1	14 (11%)	35 (24%)	
Histology			0.5
idc/ilc	118 (97%)	138 (95%)	
Other	4 (3.3%)	7 (4.8%)	
Grade of Differentiation (Pre-8th)			0.7
2	13 (11%)	16 (11%)	

3	98 (80%)	120 (83%)	
9	11 (9.0%)	9 (6.2%)	
High Grade Disease Status			0.6
0	11 (9.0%)	16 (11%)	
1	111 (91%)	129 (89%)	
LVI Status			0.7
0	79 (65%)	90 (62%)	
1	43 (35%)	55 (38%)	
HER2/neu (P/N/UK)			0.9
1	83 (68%)	100 (69%)	
2	12 (9.8%)	16 (11%)	
3	27 (22%)	29 (20%)	
HER2/neu (FISH)			0.2
1	75 (61%)	82 (57%)	
2	8 (6.6%)	6 (4.1%)	
3	14 (11%)	30 (21%)	
4	25 (20%)	27 (19%)	

HER2 Status			0.8
1	90 (74%)	109 (75%)	
2	32 (26%)	36 (25%)	
First Treatment Modality			<0.001
1	114 (93%)	107 (74%)	
2	8 (6.6%)	38 (26%)	
Attempt at Lumpectomy			0.8
0	53 (43%)	61 (42%)	
1	69 (57%)	84 (58%)	
Mastectomy Surgery			0.4
0	54 (44%)	71 (49%)	
1	68 (56%)	74 (51%)	
CPM			0.025
0	90 (74%)	123 (85%)	
1	32 (26%)	22 (15%)	
SLN Biopsy Status			0.063
0	19 (16%)	36 (25%)	

1	103 (84%)	109 (75%)	
ALND			0.5
0	57 (47%)	62 (43%)	
1	65 (53%)	83 (57%)	
Clinical T Stage			0.028
1	36 (30%)	27 (19%)	
2	80 (66%)	101 (70%)	
3	5 (4.1%)	8 (5.5%)	
4	1 (0.8%)	9 (6.2%)	
Clinical N stage			0.013
0	103 (84%)	104 (72%)	
1	19 (16%)	41 (28%)	
Primary pT Stage			0.11
T1	32 (26%)	23 (16%)	
T2	85 (70%)	114 (79%)	
T3/4	5 (4.1%)	8 (5.5%)	
Primary pN Stage			0.7

N0	64 (52%)	82 (57%)	
N1	45 (37%)	47 (32%)	
N2/3	13 (11%)	16 (11%)	
Neoadjuvant Chemotherapy Status			<0.001
0	113 (93%)	107 (74%)	
1	9 (7.4%)	38 (26%)	
Postop/Adjuvant CTX			0.069
0	11 (9.0%)	24 (17%)	
1	111 (91%)	121 (83%)	
Postop/Adjuvant XRT			0.7
None	34 (28%)	37 (26%)	
Postop/Adjuvant	88 (72%)	108 (74%)	
Adjuvant Endocrine Therapy Status			0.6
0	109 (89%)	132 (91%)	
1	13 (11%)	13 (9.0%)	
Genetic Testing			<0.001

0	66 (54%)	107 (74%)	
1	56 (46%)	38 (26%)	
Survival Time	2,935 (1,637)	480 (293)	<0.001

[1]n (%); Mean (SD)

[2]Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's exact test

Table 3. Cox PH Results (Backward Selection)

Characteristic	HR[1]	95% CI[1]	p-value
MRI Screen-Detected			
0	—	—	
1	0.00	0.00, Inf	>0.9
Index Tumor Status			
1	—	—	
2	8.28	0.62, 111	0.11
Pat Contralateral			
0	—	—	
1	0.08	0.00, 1.29	0.074
T Class - Clinical			

T1/T2	—	—	
T3/T4	1.23	0.18, 8.19	0.8
T Class - Pathologic			
1	—	—	
2	0.59	0.22, 1.58	0.3
3	0.70	0.15, 3.25	0.7
4	0.00	0.00, Inf	>0.9
Pathologic Sttage Group			
2	—	—	
3	1.34	0.47, 3.82	0.6
Reason for No Radiation			
Not recommended	—	—	
Recommended	1.67	0.30, 9.27	0.6
Radiation/Surgery Sequence			
1	—	—	
2	1.40	0.31, 6.37	0.7
High Grade Disease Status			

0	—	—	
1	5.66	0.98, 32.7	0.053
Clinical Stage Group			
1	—	—	
2	4.35	0.72, 26.4	0.11
3	2.24	0.12, 42.7	0.6
4	1.06	0.04, 29.9	>0.9
HER2/neu (FISH)			
1	—	—	
2	1.47	0.26, 8.46	0.7
3	2.20	0.64, 7.52	0.2
4	3.18	1.00, 10.1	0.050
HER2 Status			
1	—	—	
2	0.46	0.13, 1.60	0.2
First Treatment Modality			
1	—	—	

2	6.35	2.12, 19.0	<0.001
SLN Biopsy			
0	—	—	
1	2.10	0.71, 6.24	0.2
ALND			
0	—	—	
1	1.69	0.62, 4.63	0.3
Clinical N Stage			
0	—	—	
1	3.03	1.14, 8.02	0.026
Postop/Adjuvant CTX			
0	—	—	
1	0.72	0.23, 2.23	0.6
Postop/Adjuvant XRT			
None	—	—	
Postop/Adjuvant	0.13	0.03, 0.58	0.008
Age at Diagnosis	1.02	0.99, 1.06	0.2

Tumor Size - By Pathology	1.10	0.98, 1.24	0.11

[1]HR = Hazard Ratio, CI = Confidence Interval

Table 4. Cox PH Results (Elastic Net)

Characteristic	HR[1]	95% CI[1]	p-value
Clinical Stage Group			
1	—	—	
2	2.89	0.65, 12.9	0.2
3	1.77	0.23, 13.3	0.6
4	1.39	0.10, 20.1	0.8
T Class - Pathologic			
1	—	—	
2	1.05	0.48, 2.27	>0.9
3	1.77	0.58, 5.41	0.3
4	0.00	0.00, Inf	>0.9
Reason for No Radiation			
Not Recommended	—	—	
Recommended	0.83	0.38, 1.83	0.6

High Grade Disease Status			
0	—	—	
1	3.12	0.71, 13.7	0.13
First Treatment Modality			
1	—	—	
2	3.46	1.49, 8.00	0.004
Clinical N Stage			
0	—	—	
1	2.29	1.06, 4.94	0.035

[1]HR = Hazard Ratio, CI = Confidence Interval

Table 5. Important Predictors from RSF

var_name	importance
age_diagnosis	0.10
neoadjuvant_chemo1	0.08
clinical_n_stage1	0.08
first_treatment_modality2	0.07
tumor_size_path	0.06
regional_lymph_nodes_exami	0.06

path_stage_group3	0.05
her2_neu_fish3	0.04
postop_adjuvant_ctx1	0.03
postop_adjuvant_xrtpostop_adjuvant	0.03
t_class_clinicalT3/T4	0.02
surgical_procedure_of_primtotal_masterctomy	0.02
her2_neu_fish4	0.02
lvi1	0.02
spanish_originunknown	0.02
sln_biopsy1	0.02
index_tumor_status2	0.02
mammo_screen1	0.02
radiation_surgery_sequence2	0.02
t_class_pathologic2	0.01
mastectomy_surgery1	0.01
genetic_testing1	0.01
pat_contralateral1	0.01
attempt_lumpectomy1	0.01
matatases_dx1	0.01
primary_pn_stageN2/3	0.01

alnd1	0.01
t_class_pathologic3	0.01
primary_pn_stageN1	0.01
reason_for_no_radiationrecommanded	0.01
clinical_t_stage2	0.01
raceWA	0.01
regional_ln_surgeryregional_ln_surgery	0.01
clin_stage_group3	0.01
her2_neu_p_n_uk2	0.01
adju_endo_therapy1	0.01
grade_of_differentiation_p3	0.01
hormone_therapy1	0.01
primary_pt_stageT2	0.00
primary_pt_stageT3/4	0.00
cpm1	0.00
raceUK	0.00
her2_neu_p_n_uk3	0.00
her2_status2	0.00
clin_stage_group2	0.00
high_grade_disease1	0.00

clinical_t_stage3	0.00
her2_neu_fish2	0.00
clinical_t_stage4	0.00
grade_of_differentiation_p9	0.00
surgical_procedure_of_primOthers	0.00
spanish_originspanish	0.00
past_ipsilateral1	0.00
raceAsian_Pacific	0.00
raceOther	0.00
clin_stage_group4	0.00
(Intercept)	0.00
mammo_occult1	0.00
t_class_pathologic4	0.00
histologyothers	0.00

Figure 1. Tuning Paramter Selection (Elastic Net Regression)

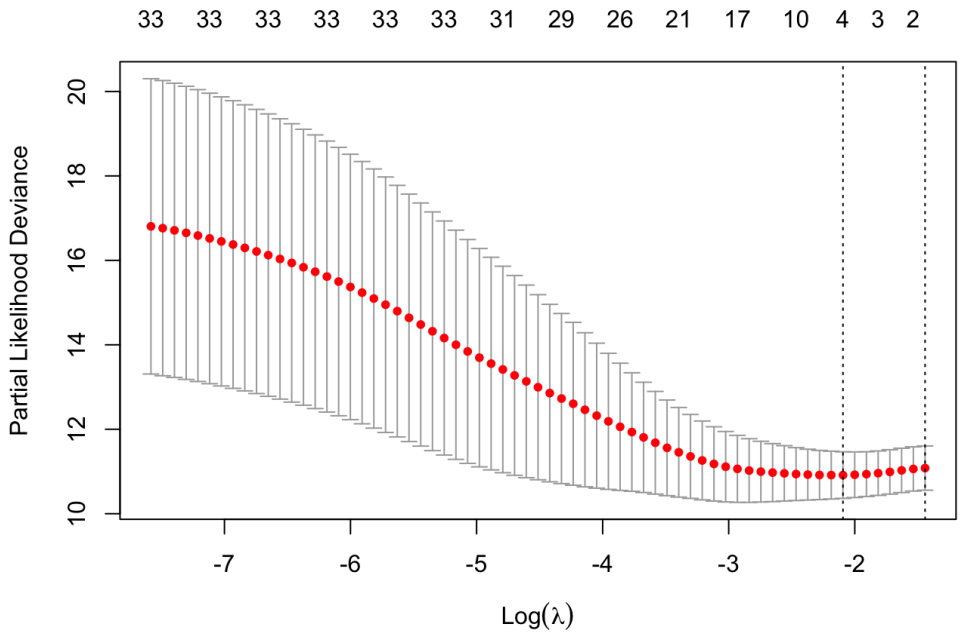


Figure 2. ROC for Backwards Selection Model and Elastic Net Selection Model