# P8130 Final Project: Predicting Breast Cancer Survival

Guadalupe Antonio Lopez, Gustavo Garcia-Franceschini, Derek Lamb

UNI's: GA2612, GEG2145, DRL2168

## Abstract

Studies have shown that there are many risk factor combinations for breast cancer. Fortunately, in recent years, breast cancer survival rates have increased and number of deaths have decreased. Nevertheless, it is important to explore the most significant breast cancer risk factors to identify at-risk groups for further research. To do so, we used a dataset containing demographic and risk factors information, and survival status by patients. After cleaning and exploring the data, we dropped highly correlated variables, combined others, and performed transformations. We then built the full model, containing an interaction between estrogen and progesterone status, and selected the best model using Akaike Information Criterion (AIC). Overall, our final model had a good ROC-AUC and Brier score values, enabling it to effectively predict likelihood of death in breast cancer patients.

## Introduction

Breast cancer occurs due to abnormal cell growths in breast tissue. Although it is most often found in females, 1 out of every 100 diagnosed patients in the US is a male. Other breast cancer risk factors include increased age, family history or personal history of breast cancer, radiation exposure, obesity, alcohol use, and more. Research suggests that postmenopausal hormone therapy is a risk factor due the combination of estrogen and progesterone used to treat signs and symptoms of menopause.

Additionally, the patient's breast cancer stage is important to consider when determining the severity of the cancer and how to treat it. The American Joint Committee on Cancer (AJCC) TNM system is the most common, and contains clinical and pathologic systems. The pathologic stage (also known as the surgical stage) is determined by examining the tissue removed during surgery, while the clinical stage is based on results of a physical exam, biopsy, and imaging tests. Nevertheless, both systems are composed of the size of the tumor, the spread to nearby lymph nodes and/or to distant sites, their estrogen and/or progesterone receptor status, the grade of the cancer, and if the cancer makes too much of HER2 protein.

Although breast cancer survival rates have increased and number of deaths have decreased recently, it is important to explore the risk factors of breast cancer. Understanding these risk factors allow for the identification of still high-risk patient populations and the tailoring of cancer treatment to the patient in the changing paradigm of personalized medicine. For this project, we will investigate the odds of breast cancer survival given most of the risk factors previously mentioned.

## Methods

**Data source** We obtained a deidentified set containing data on 4024 breast cancer patients. this dataset contains both demographic information, such as patient age, race, and marital status; clinical information such as tumor stage, tumor size, cancer hormone receptor status (progesterone and estrogen), regional node positive, and regional node examined; and outcome information: the number of months the patient had survived prior to study conclusion, and their alive/dead status at the end of the study.

**Data cleaning** We combined the regional node positive and regional node examined variables into a "regional node proportion positive" variable. This variable, but neither the node positive nor node examined variables were in the model. Further, we decided to discard the T stage and N stage variables, as they captured information already contained in the AJCC 6th stage variable. We also excluded the grade variable, as it captured the same clinical information as the differentiate variable. Due to the skewness in the distribution of the tumor size, we applied a square root transformation to that variable (**Supplemental Figure 1**). We also added a main_stage variable, which groups all the stages in 6th_stage, i.e. that "IIIA" and "IIIC" are under factor level "III", and added an indicator variable that tells us if the patient is "White" or not, essentially grouping "Black" and "Other" together. Both variables were created in case the reduction of factor levels makes the fit better.

**Model construction** We decided to use logistic regression model to estimate the risk of patient death within the followup window. Formally, we assumed that for an individual with probability $p$ to die after receiving a breast cancer diagnosis, the log-odds of $p$ was linear, i.e. $logit(p) = \mathbf{X}\beta + \epsilon$, where $\mathbf{X}$ is the n x p design matrix, and $\beta$ and $\epsilon$ are vectors in $\mathbb{R}^p$.

In addition to the covariates, we included the interaction between estrogen status and progesterone status given that we found in our background research that having both "positive" increase the chances of breast cancer.

**Model selection** We used a criterion-based method, utilizing Akaike Information Criterion (AIC) to assess the performance of our models.

**Model validation** We performed 10-fold cross-validation to assess the performance of our model. Each observation in a dataset will be in 1 of 10 folds such that it gets used as training data 9 times, and as test data once. Because this dataset has over 4,000 subjects, each test set will include approximatley 400 subjects. The predictions are then saved such that each row has an out-of-sample prediction that can be compared to the real value.

## Results

**Exploratory Data Analysis** **Table 1** shows summary statistics for demographic variables, along with some variables related to breast cancer. They are split by whether they survived or not, with the first column being the summary statistics for the entire dataset. We noticed that although individuals had similar ages despite their status, Black individuals and widows died at a disproportional rate. Those that died also had higher root_tumor_size and regional_prop, which are variables related to cancer where higher values are more alarming. Summary statistics for all variables can be found in **Supplemental Table 1**.

**Table 1.** Baseline characteristics

| **Characteristic** | **Overall**, N = 4,024 | **Alive**, N = 3,408 | **Dead**, N = 616 |
|---|---|---|---|
| age | 54 (9) | 54 (9) | 55 (10) |
| race | | | |
| White | 3,413 / 4,024 (85%) | 2,903 / 3,408 (85%) | 510 / 616 (83%) |
| Black | 291 / 4,024 (7.2%) | 218 / 3,408 (6.4%) | 73 / 616 (12%) |
| Other | 320 / 4,024 (8.0%) | 287 / 3,408 (8.4%) | 33 / 616 (5.4%) |
| marital_status | | | |
| Married | 2,643 / 4,024 (66%) | 2,285 / 3,408 (67%) | 358 / 616 (58%) |
| Divorced | 486 / 4,024 (12%) | 396 / 3,408 (12%) | 90 / 616 (15%) |
| Single | 615 / 4,024 (15%) | 511 / 3,408 (15%) | 104 / 616 (17%) |
| Widowed | 235 / 4,024 (5.8%) | 186 / 3,408 (5.5%) | 49 / 616 (8.0%) |

| **Characteristic** | **Overall**, N = 4,024 | **Alive**, N = 3,408 | **Dead**, N = 616 |
|---|---|---|---|
| Separated | 45 / 4,024 (1.1%) | 30 / 3,408 (0.9%) | 15 / 616 (2.4%) |
| root_tumor_size | 5.24 (1.73) | 5.14 (1.69) | 5.81 (1.85) |
| regional_prop | 0.33 (0.29) | 0.30 (0.27) | 0.49 (0.33) |

**Model construction and selection** We used a logistic model coupled with criterion-based stepwise regression to determine which variables were useful in predicting the risk of death in breast cancer patients. The variables that were identified as important were age, race, marital status, AJCC 6th stage, differentiate, estrogen status, progesterone status, tumor size, and regional node positive proportion. Some variables that were not identified as important by the model were whether the tumor was Stage A and the interaction between estrogen and progesterone status. For a list of model coefficients see **Supplemental Table 2**.

**Diagnostics** We constructed a receiver operating characteristic curve (ROC) for our model on in-sample data, and used the area under the curve (AUC) to measure the model's discriminatory capability. This approach compares the model's specificity and sensitivity to get a score bounded by 0 and 1, where 1 is best. This model's ROC-AUC is 0.754.

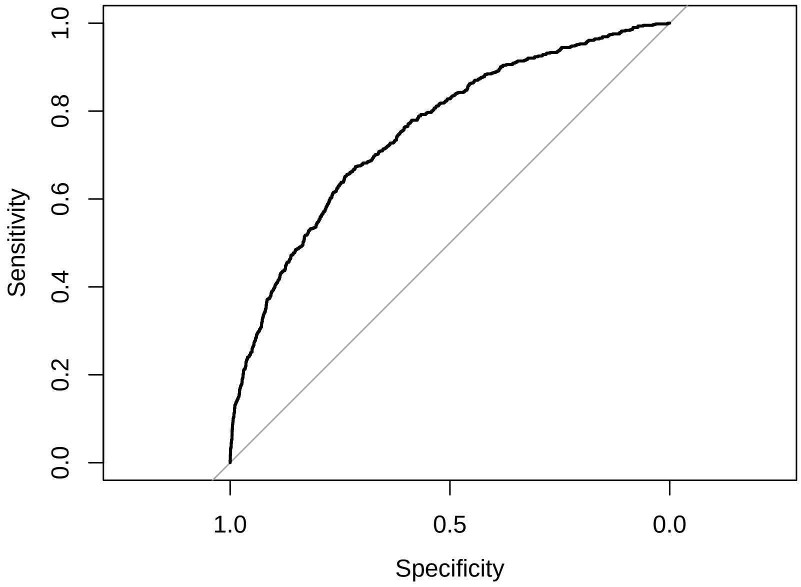

**Figure 1.** Receiver operating characteristic curve for the optimal logistic regression model.

We calculated Brier scores to assess the optimal model's performance. The Brier score measures the accuracy of probabilistic predictions, where a score of 0 indicates perfect accuracy. The optimal model's Brier score was 0.112. This suggests that our final model has good probabilistic prediction accuracy.

To further assess the performance of our model, we constructed a separation plot, a graphical method for determining concordant and discordant predictions (**Figure 2**). This separation plot shows that the model is able to reasonably effectively distinguish the patients that survived from those that died, though the low value of the predicted probability across the sample shows that it gives low probabilities of dying to all subjects.

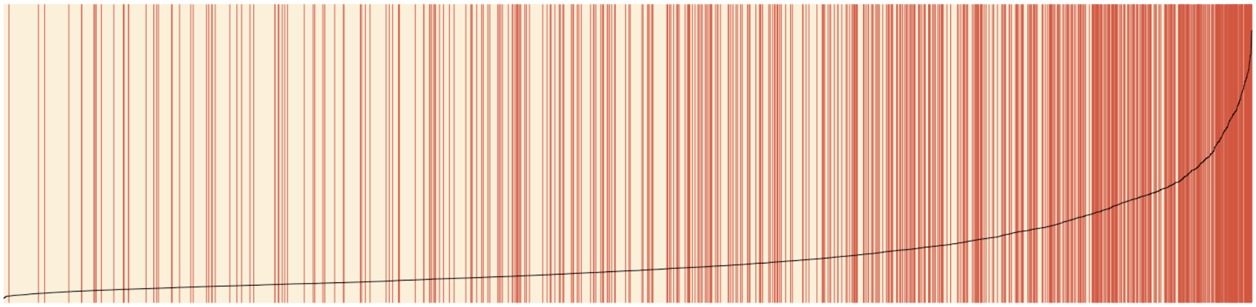

**Figure 2.** Separation plot of model. Values are stripes, arranged in increasing predicted probability of death. The stripes are colored yellow if the patient survived, and red if they died. The black line indicates the predicted probability of death.

**Model validation** This table shows the confusion matrix made from our 10 fold cross-validation, such that the predicted values are out-of-sample predictors. Note that cells 1 and 4 (if read from left to right, and top-down) are individuals for which their status prediction was correctly predicted, and cells 2 and 3 are individuals for which their status wasn't correctly predicted.

**Table 1.** Confusion matrix of the model

| | Predicted Died | Predicted Survived |
|---|---:|---:|
| Actually Died | 3355 | 53 |
| Actually Survived | 536 | 80 |

**Model performance by race** Lastly, model performance by race was explored using ROC AUC and Brier score values. The values by race are presented in **Table 2**. The highest AUC was obtained for white patients, while the lowest Brier score was obtained for the Other racial group. Separation plots were also constructed stratified on racial groups (**Figure S2**). These plots showed good performance for White subjects, but poorer performance for subjects of Black and Other race.

**Table 2.** Model performance by race

| race | roc_auc | brier_score |
|---|---|---|
| White | 0.760 | 0.109 |
| Black | 0.704 | 0.171 |
| Other | 0.650 | 0.088 |

## Discussion

We constructed a model to predict the survival of breast cancer patients, and optimized it using a criterion based-selection approach. Most of the covariates in the dataset were identified by AIC as being useful predictors, with the exception of Stage A. Despite the combination of progesterone (PR+) and estrogen (ER+) receptors in breast cancer being of note clinically, our model did not find the interaction term between these variables to be worth including. This suggests that the odds ratio for risk of death in ER+ patients with PR+ is not significantly different from those who are PR-, and vice versa.

The most significant predictor of death was the regional node positive proportion variable; an individual with 100% positive regional nodes has 3.5 times the odds of death (CI: 2.4, 5.0) compared with an individual who had no positive regional nodes. Other highly significant predictors were cancer stage (specifically whether the patient was IIIC), whether the tumor was well differentiated (i.e. healthy cells), hormone receptors on the tumor (ER+ and PR+), and age.

Our model was effective at classifying patients who survived, but was less effective at classifying patients who died when using a $\hat{p} = 0.5$ as the cutoff. The separation plot we constructed suggests that a lower cutoff, such as 0.35 may be more appropriate, and that further tuning may improve model performance.

Conflicting conclusions occurred while assessing model performance by race. According to the ROC AUC values, the model performed better for White patients, then Black patients, then 'Other' race patients. However, the Brier scores by race suggests that the model most accurately predicted probabilities for 'Other' race patients, followed by White and Black patients. None of these scores were alarming to the extent we thought the model had horrible performance on one of the race groups, but any possible improvements would likely require more data or a more flexible model.

From our cross-validation confusion matrix, we can see that the model is very good at correctly predicting that somebody survived, and rarely predicts the individual died when it didn't. However, the model makes a lot of errors when trying to predict if the individual survived: more than 90% of the total error happens when the model predicts the individual died, and it actually survived. This reflects the fact that the model prefers to predict 0 as a result of the status variable being 0 inflated.

Overall, our model provides useful insight into the risk factors for death in breast cancer patients. Future research may expand this project even further by using a dataset with more non-white participants, thus improving prediction across racial strata. Other models, like random forest, could lead to prediction accuracy improvement. This further exploration is important to improving equity in healthcare outcomes, and is in line with our goal of identifying vulnerable patient subgroups for further research.

## Author Contributions

Guadalupe wrote the abstract and introduction and performed model diagnostics (Brier scores). Gustavo performed exploratory data analysis, model performance, and model diagnostics (ROC curves). Derek did the data cleaning, model construction, and diagnostics (separation plot). All authors contributed to the writing of the results and the discussion.

## References

American Cancer Society. Breast Cancer Stages. Atlanta, GA: American Cancer Society, Inc. Available from: https://www.cancer.org/cancer/types/breast-cancer/

Division of Cancer Prevention and Control, Centers for Disease Control and Prevention. Breast Cancer in Men. Atlanta, GA: Centers for Disease Control; 2023 July 25. Available from: https://www.cdc.gov/cancer/breast/men/

Greenhill B, Ward MD, Sacks A. The separation plot: A new visual method for evaluating the fit of binary models. American Journal of Political Science. 2011;55(4):991–1002. doi:[10.1111/j.1540-5907.2011.00525.x](doi:%5B10.1111/j.1540-5907.2011.00525.x){.uri}

Mayo Clinic. Breast Cancer. Rochester, MN: Mayo Foundation for Medical Education and Research. Available from: https://www.mayoclinic.org/diseases-conditions/breast-cancer

Stoica P, Selen Y. Model-order selection: a review of information criterion rules. IEEE Signal Processing Magazine. 2004;21(4):36–47. doi:10.1109/msp.2004.1311138

Young Survival Coalition. Breast Cancer 101. New York, NY: Young Survival Coalition. Available from: https://youngsurvival.org/breast-cancer

**Software** The aforementioned analyses were carried out using R 4.3.1 and RStudio Version 2023.06.2+561.

## Supplemental Information

**Table S1. Summary Statistics for all variables**

| **Characteristic** | **Overall**, N = 4,024 | **Alive**, N = 3,408 | **Dead**, N = 616 |
|---|---|---|---|
| age | 54 (9) | 54 (9) | 55 (10) |
| race | | | |
| White | 3,413 / 4,024 (85%) | 2,903 / 3,408 (85%) | 510 / 616 (83%) |
| Black | 291 / 4,024 (7.2%) | 218 / 3,408 (6.4%) | 73 / 616 (12%) |
| Other | 320 / 4,024 (8.0%) | 287 / 3,408 (8.4%) | 33 / 616 (5.4%) |
| marital_status | | | |
| Married | 2,643 / 4,024 (66%) | 2,285 / 3,408 (67%) | 358 / 616 (58%) |
| Divorced | 486 / 4,024 (12%) | 396 / 3,408 (12%) | 90 / 616 (15%) |
| Single | 615 / 4,024 (15%) | 511 / 3,408 (15%) | 104 / 616 (17%) |
| Widowed | 235 / 4,024 (5.8%) | 186 / 3,408 (5.5%) | 49 / 616 (8.0%) |
| Separated | 45 / 4,024 (1.1%) | 30 / 3,408 (0.9%) | 15 / 616 (2.4%) |
| x6th_stage | | | |
| IIA | 1,305 / 4,024 (32%) | 1,209 / 3,408 (35%) | 96 / 616 (16%) |
| IIIA | 1,050 / 4,024 (26%) | 866 / 3,408 (25%) | 184 / 616 (30%) |
| IIIC | 472 / 4,024 (12%) | 291 / 3,408 (8.5%) | 181 / 616 (29%) |
| IIB | 1,130 / 4,024 (28%) | 995 / 3,408 (29%) | 135 / 616 (22%) |
| IIIB | 67 / 4,024 (1.7%) | 47 / 3,408 (1.4%) | 20 / 616 (3.2%) |
| differentiate | | | |
| Poorly differentiated | 1,111 / 4,024 (28%) | 848 / 3,408 (25%) | 263 / 616 (43%) |
| Moderately differentiated | 2,351 / 4,024 (58%) | 2,046 / 3,408 (60%) | 305 / 616 (50%) |
| Well differentiated | 543 / 4,024 (13%) | 504 / 3,408 (15%) | 39 / 616 (6.3%) |
| Undifferentiated | 19 / 4,024 (0.5%) | 10 / 3,408 (0.3%) | 9 / 616 (1.5%) |
| a_stage | | | |
| Regional | 3,932 / 4,024 (98%) | 3,351 / 3,408 (98%) | 581 / 616 (94%) |
| Distant | 92 / 4,024 (2.3%) | 57 / 3,408 (1.7%) | 35 / 616 (5.7%) |
| estrogen_status | | | |
| Positive | 3,755 / 4,024 (93%) | 3,247 / 3,408 (95%) | 508 / 616 (82%) |
| Negative | 269 / 4,024 (6.7%) | 161 / 3,408 (4.7%) | 108 / 616 (18%) |

| **Characteristic** | **Overall**, N = 4,024 | **Alive**, N = 3,408 | **Dead**, N = 616 |
|---|---|---|---|
| progesterone_status | | | |
| Positive | 3,326 / 4,024 (83%) | 2,914 / 3,408 (86%) | 412 / 616 (67%) |
| Negative | 698 / 4,024 (17%) | 494 / 3,408 (14%) | 204 / 616 (33%) |
| root_tumor_size | 5.24 (1.73) | 5.14 (1.69) | 5.81 (1.85) |
| regional_prop | 0.33 (0.29) | 0.30 (0.27) | 0.49 (0.33) |
| main_stage | | | |
| II | 2,435 / 4,024 (61%) | 2,204 / 3,408 (65%) | 231 / 616 (38%) |
| III | 1,589 / 4,024 (39%) | 1,204 / 3,408 (35%) | 385 / 616 (63%) |
| white | | | |
| POC | 611 / 4,024 (15%) | 505 / 3,408 (15%) | 106 / 616 (17%) |
| White | 3,413 / 4,024 (85%) | 2,903 / 3,408 (85%) | 510 / 616 (83%) |
| cv | 6 (3) | 6 (3) | 6 (3) |
| predicted_prob | 0.15 (0.13) | 0.13 (0.11) | 0.27 (0.19) |
| resid | 2.15 (0.89) | 2.14 (0.85) | 2.20 (1.07) |
| pred | -2.00 (0.95) | -2.14 (0.85) | -1.20 (1.07) |
| p_hat | 0.15 (0.13) | 0.13 (0.11) | 0.27 (0.19) |

**Table S2. Model Coefficients**

| term | estimate | std.error | adjusted_odds_ratio | p.value | statistic |
|---|---|---|---|---|---|
| age | 0.023 | 0.006 | 1.02 (1.01, 1.03) | 0.000 | 4.180 |
| raceBlack | 0.506 | 0.162 | 1.66 (1.21, 2.28) | 0.002 | 3.129 |
| raceOther | -0.431 | 0.202 | 0.65 (0.44, 0.97) | 0.033 | -2.132 |
| marital_statusDivorced | 0.222 | 0.141 | 1.25 (0.95, 1.65) | 0.115 | 1.575 |
| marital_statusSingle | 0.153 | 0.134 | 1.17 (0.9, 1.52) | 0.252 | 1.146 |
| marital_statusWidowed | 0.225 | 0.192 | 1.25 (0.86, 1.83) | 0.242 | 1.171 |
| marital_statusSeparated | 0.847 | 0.366 | 2.33 (1.14, 4.78) | 0.021 | 2.316 |
| x6th_stageIIIA | 0.563 | 0.163 | 1.76 (1.28, 2.42) | 0.001 | 3.460 |
| x6th_stageIIIC | 1.064 | 0.193 | 2.9 (1.98, 4.23) | 0.000 | 5.505 |
| x6th_stageIIB | 0.413 | 0.154 | 1.51 (1.12, 2.05) | 0.007 | 2.676 |
| x6th_stageIIIB | 1.141 | 0.322 | 3.13 (1.67, 5.88) | 0.000 | 3.546 |

| term | estimate | std.error | adjusted_odds_ratio | p.value | statistic |
|---|---|---|---|---|---|
| differentiateModerately differentiated | -0.389 | 0.104 | 0.68 (0.55, 0.83) | 0.000 | -3.726 |
| differentiateWell differentiated | -0.919 | 0.192 | 0.4 (0.27, 0.58) | 0.000 | -4.776 |
| differentiateUndifferentiated | 0.961 | 0.529 | 2.61 (0.93, 7.37) | 0.069 | 1.816 |
| estrogen_statusNegative | 0.738 | 0.177 | 2.09 (1.48, 2.96) | 0.000 | 4.169 |
| progesterone_statusNegative | 0.571 | 0.127 | 1.77 (1.38, 2.27) | 0.000 | 4.485 |
| root_tumor_size | 0.048 | 0.031 | 1.05 (0.99, 1.12) | 0.128 | 1.523 |
| regional_prop | 1.237 | 0.185 | 3.45 (2.4, 4.95) | 0.000 | 6.686 |

**Supplemental figures**

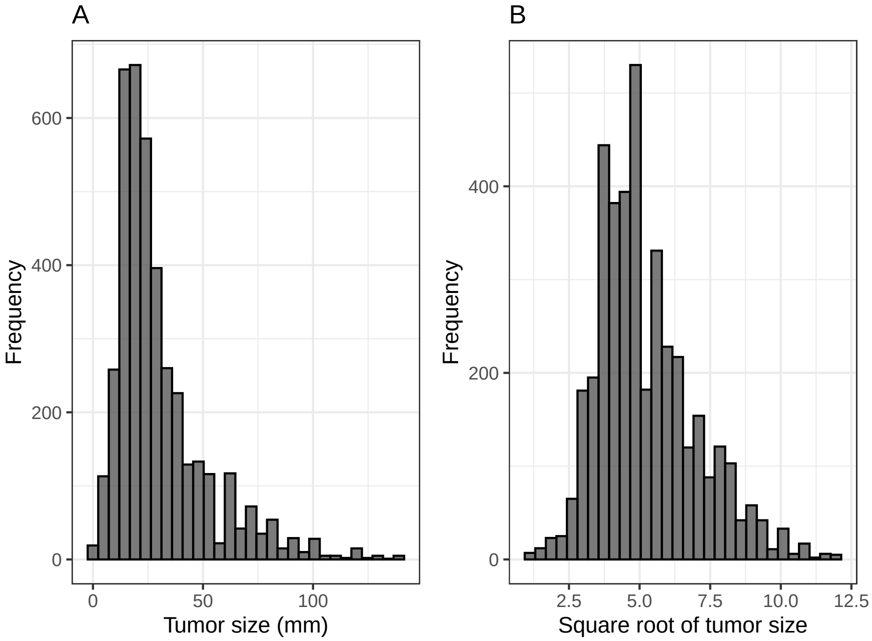

**Figure S1.** Transformation of tumor size variable. (A) Before transformation. (B) After square root transformation.

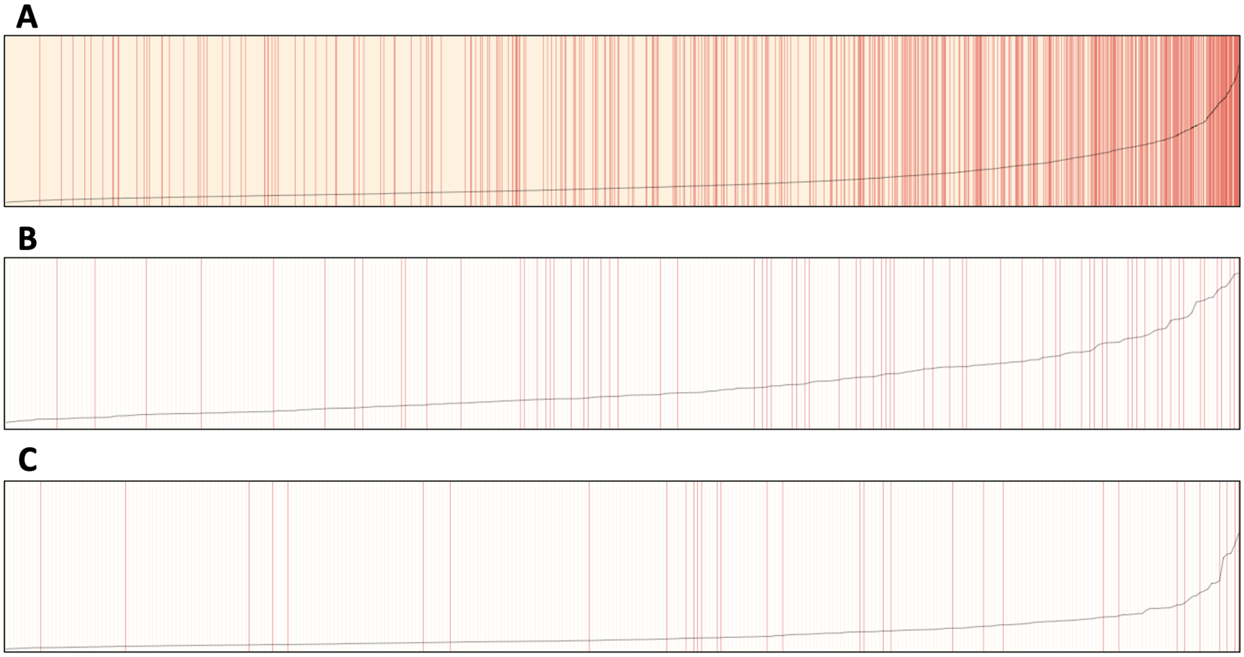

**Figure S2.** Separation plots by race. Values are stripes, arranged in increasing predicted probability of death in (A) White, (B) Black, and (C) Other race patients. The stripes are colored yellow if the patient survived, and red if they died. The black line indicates the predicted probability of death.