# Untitled

Some authors

## Abstract

## Introduction

Background info/research on breast cancer patients:

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, increase in age, family history or personal history of breast cancer, radiation exposure, obesity, alcohol use, among many 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 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 protien.

Stages are most commonly based on

Most recently, breast cancer survival rates have increased and number of deaths decreased.

## 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, hormone therapies (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 decided to combine 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**).

**Model construction** We decided to use logistic regression model to estimate the risk of patient death within the followup window. Formally, we assumed that an for an individual, with probability $p$ to die after receiving a breast cancer diagnosis, the log-odds of $p$ was linear, i.e.

$logit(p_i) = \mathbf{X}\beta + \epsilon_i$

Where $\boldsymbol{X}$ is the n x p design matrix, and $\beta$ is a vector in $\mathbb{R}^p$.

In addition to the covariates we included the following interaction terms [].

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

**Model validation** We performed [size] cross-validation to assess the performance of our model.

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

## Results

**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. The variables that were not identified as important by the model were whether the tumor was Stage A and the interaction between estrogen and progresterone status. For a list of model coefficients see **Supplemental Table 1**.

**Diagnostics** ROC, Brier score

Separation plots

**Model performance by race** Make a table with ROC AUC and Brier scores by race. Discuss

## Discussion

## Author Contributions

## References

## Appendices

a. tumor size transformation

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

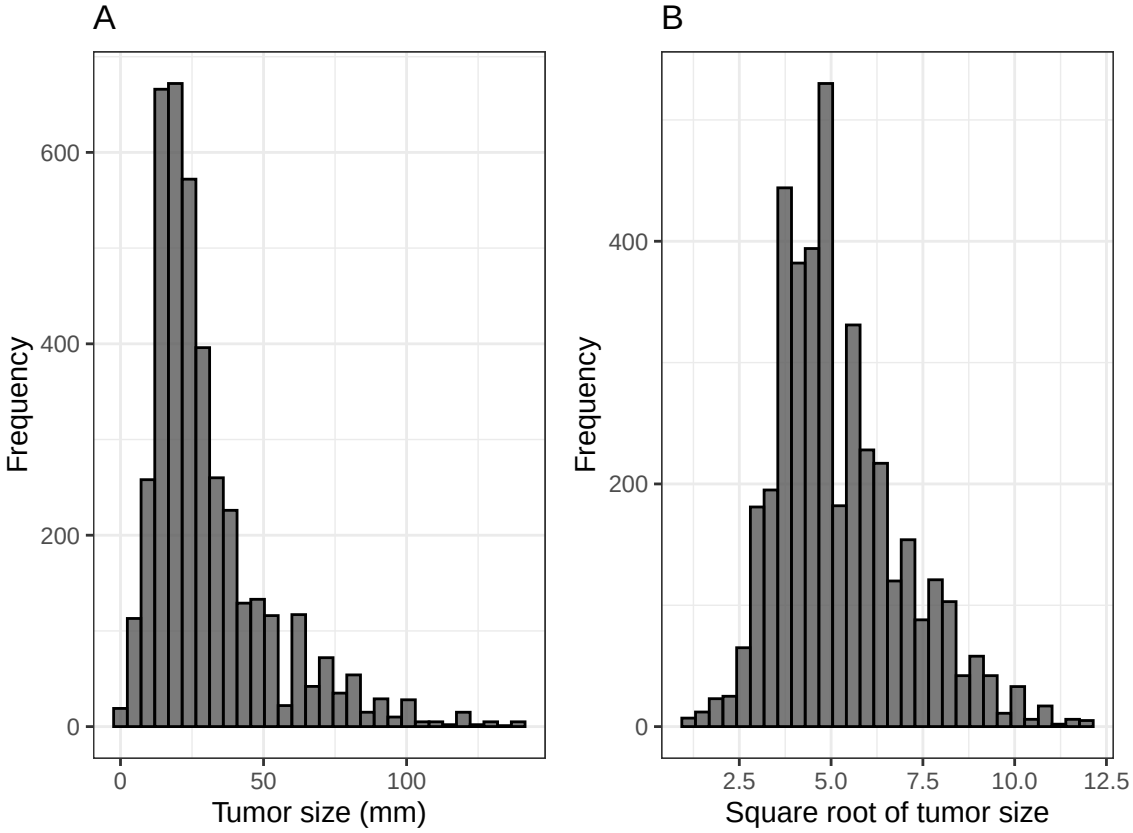


**Figure S1.** Transformation of tumor size variable. (A) Before transformation. (B) After square root transformation. b. model coefficients

**Table S1. Model Coefficients**

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -4.197 | 0.369 | -11.382 | 0.000 |
| age | 0.023 | 0.006 | 4.180 | 0.000 |
| raceBlack | 0.506 | 0.162 | 3.129 | 0.002 |
| raceOther | -0.431 | 0.202 | -2.132 | 0.033 |
| marital_statusDivorced | 0.222 | 0.141 | 1.575 | 0.115 |
| marital_statusSingle | 0.153 | 0.134 | 1.146 | 0.252 |

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| marital_statusWidowed | 0.225 | 0.192 | 1.171 | 0.242 |
| marital_statusSeparated | 0.847 | 0.366 | 2.316 | 0.021 |
| x6th_stageIIIA | 0.563 | 0.163 | 3.460 | 0.001 |
| x6th_stageIIIC | 1.064 | 0.193 | 5.505 | 0.000 |
| x6th_stageIIB | 0.413 | 0.154 | 2.676 | 0.007 |
| x6th_stageIIIB | 1.141 | 0.322 | 3.546 | 0.000 |
| differentiateModerately differentiated | -0.389 | 0.104 | -3.726 | 0.000 |
| differentiateWell differentiated | -0.919 | 0.192 | -4.776 | 0.000 |
| differentiateUndifferentiated | 0.961 | 0.529 | 1.816 | 0.069 |
| estrogen_statusNegative | 0.738 | 0.177 | 4.169 | 0.000 |
| progesterone_statusNegative | 0.571 | 0.127 | 4.485 | 0.000 |
| root_tumor_size | 0.048 | 0.031 | 1.523 | 0.128 |
| regional_prop | 1.237 | 0.185 | 6.686 | 0.000 |

b. ROC by race +/- sep plots?