Untitled

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. An interesting risk factor: postmenopausal hormone therapy (combines estrogen and progesterone to treat signs and symptoms of menopause).

Most recently, breast cancer survival rates have increase 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 patient information as the differentiate variable.

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 m x n design matrix, and β is a vector in $\mathbb{R}^n$.

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.