Determining the Most Effective Treatment for Breast Cancer Types

Independent Research Project

Kaitlyn Ng

Prediction Question

Using Kaggle:/METABRIC_RNA_Mutation.csv, how can we predict what type of breast cancer surgery will be most effective based on the status of the ER and HER2 receptors, number of positive lymph nodes, and detailed cancer type?

Research Proposal

Introduction

Breast cancer is a genetic disease that has affected millions of women worldwide, representing around 15% of cancer cases every year (SEER). In 2021, about 43,600 people will die from breast cancer (SEER). In an effort to reduce this number, this study will be determining the most effective breast cancer surgery based on the specific patient's characteristics by utilizing information from the METABRIC database.

Literature Review

Previous work has been focused on utilizing or creating new "-omic" technology to develop a better understanding of breast cancer heterogeneity (Caswell-Jin et al., 2021, p. 88). However, current biomarkers focus specifically within one phase of tumor development: the untreated primary tumor, thus they are not fully equipped to predict the best specific treatment options for breast cancer patients because of the limited coverage of the tumor's heterogeneity (Caswell-Jin et al., 2021, p. 88-89). This study seeks to better improve the prediction of treatment for breast cancer patients through machine learning.

Research Proposal (cont.)

Purpose

This study aims to answer the question of how we can predict what type of breast cancer surgery will be most effective based on the status of the ER, HER2, and PR receptors, number of positive lymph nodes, and detailed cancer type utilizing Kaggle:/METABRIC_RNA_Mutation.csv. The basal-like subtype of breast cancer is exceptionally difficult to treat, with only chemotherapy being an effective solution (Urry et al., 2017). Thus, by analyzing the best surgery option for individuals with basal-like breast cancer, improvements can be made to increase the treatment options available.

Methodology

This study will be focused on data from the METABRIC dataset, which features genetic information from 1,980 breast cancer samples (Pereira et al., 2016). This dataset does not need cleaning as there is minimal data missing (Aparicio & Caldas, 2016). Because the dataset contains roughly 700 columns, a new version of the dataset will be created by splicing the columns to only include "age_at_diagnosis," "type_of_breast_surgery," "cancer_type_detailed," "er_status_measured_by_ihc," "her2_status_measured_by_snp6," "lymph_nodes_examined_positive," "overall_survival_months," "overall_survival," "pr_status," and "death_from_cancer" (Aparicio & Caldas, 2016). Next, groupby functions will be implemented to search for patterns, ie. is there a correlation between a specific type of surgery given to a patient and their HER2 status? Then a linear regression model will be fitted to the data in order to predict the overall survival of a patient for each surgery option using the patterns found and survival data. The surgery option that has the highest rate of survival, should be the optimal option for that specific patient.

Research Proposal (cont.)

References

Aparicio, S. Caldas, C. (2016). *The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database* [Data set]. CBioPortal. https://www.kaggle.com/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

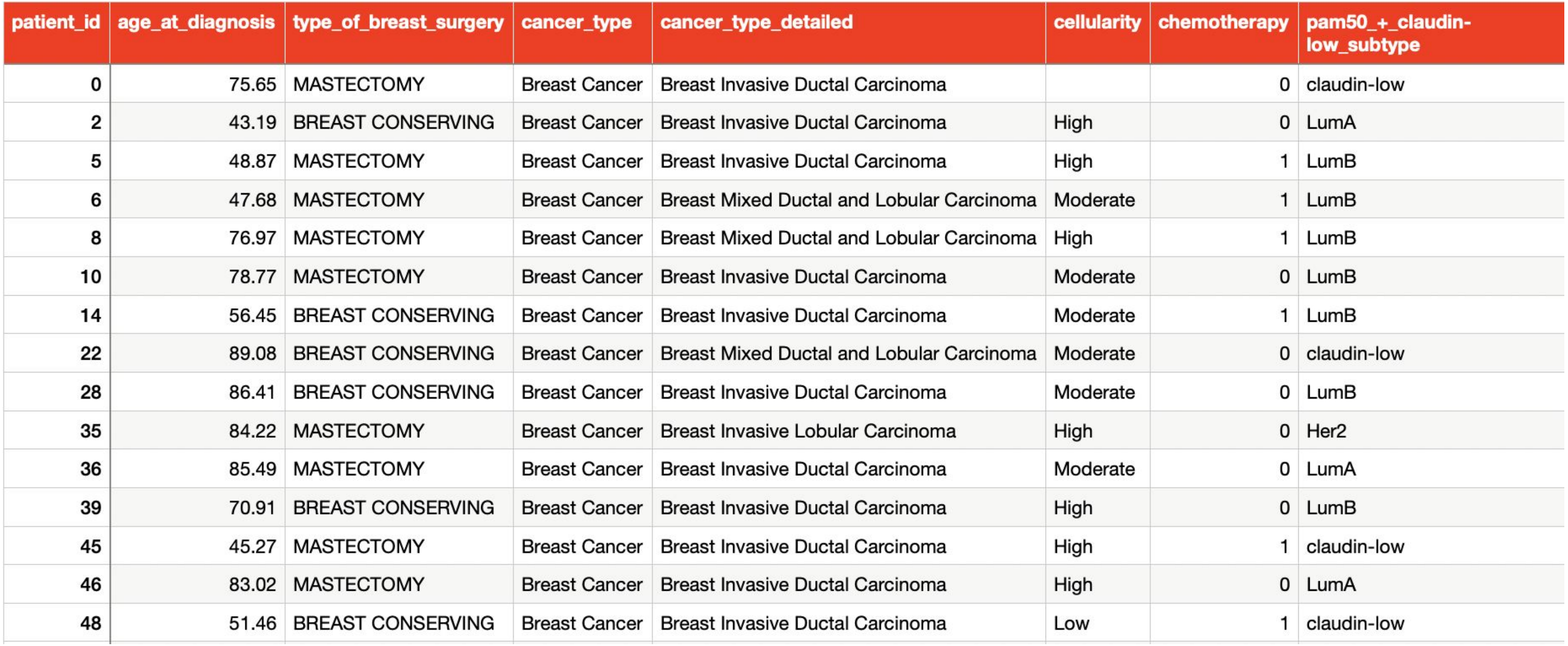
Cancer of the Breast (Female) - Cancer Stat Facts. SEER. (n.d.). https://seer.cancer.gov/statfacts/html/breast.html.

Caswell-Jin, J. L., Lorenz, C., & Curtis, C. (2021). Molecular Heterogeneity and Evolution in Breast Cancer. *Annual Review of Cancer Biology*, *5*(1), 79–94. https://doi.org/10.1146/annurev-cancerbio-060220-014137

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D. W. Y., Liu, B., Dawson, S.-J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., … Caldas, C. (2016, May 10). *The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes*. Nature News. https://www.nature.com/articles/ncomms11479.

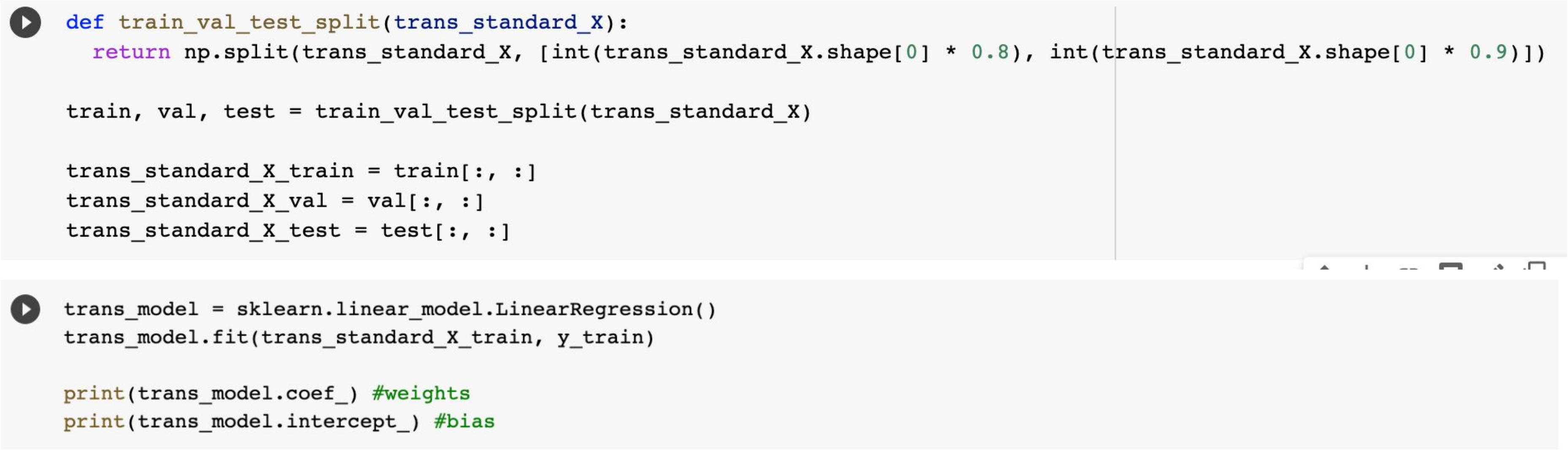
Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., & Reece, J. B. (2017). *Campbell Biology*. Pearson.

METABRIC Dataset



patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	cellularity	chemotherapy	pam50_+_claudin-low_subtype
0	75.65	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma		0	claudin-low
2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumA
5	48.87	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	1	LumB
6	47.68	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	1	LumB
8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	1	LumB
10	78.77	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0	LumB
14	56.45	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	1	LumB
22	89.08	BREAST CONSERVING	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	0	claudin-low
28	86.41	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0	LumB
35	84.22	MASTECTOMY	Breast Cancer	Breast Invasive Lobular Carcinoma	High	0	Her2
36	85.49	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0	LumA
39	70.91	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumB
45	45.27	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	1	claudin-low
46	83.02	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumA
48	51.46	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	Low	1	claudin-low

Machine Learning Python Code



```
def train_val_test_split(trans_standard_X):
  return np.split(trans_standard_X, [int(trans_standard_X.shape[0] * 0.8), int(trans_standard_X.shape[0] * 0.9)])

train, val, test = train_val_test_split(trans_standard_X)

trans_standard_X_train = train[:, :]
trans_standard_X_val = val[:, :]
trans_standard_X_test = test[:, :]
```

```
trans_model = sklearn.linear_model.LinearRegression()
trans_model.fit(trans_standard_X_train, y_train)

print(trans_model.coef_) #weights
print(trans_model.intercept_) #bias
```

Code still work in progress!

Results

Work in progress!