Predictive Modeling of Estrogen and Progesterone Hormonal Imbalance Effects on Breast Cancer Incidence and Survival Rate

Team members: Priyankitha Kandi, Kalyan Raj Chinigi, Kavya Surabhi, Haneesha Chowdary Dondepudi, Srihari Myla Venkata, Saikumar Reddy Yalagalapalli

1. Introduction

Breast cancer is one of the most common cancers affecting women worldwide accounting for over 2 million new cases annually. It is a complex, heterogeneous disease with multiple risk factors contributing to its initiation and progression. The dataset contains information about the patients' medical histories, including tumor size, lymph node status, tumor grade, and survival status. Among these factors, hormonal imbalances play a significant role in breast cancer pathogenesis. Estrogen and progesterone are key hormones that regulate the growth and development of breast tissue. However, disruptions in the delicate balance of these hormones can promote abnormal cellular proliferation and survival, leading to breast cancer development. Specifically, excessive exposure to estrogen and impaired progesterone signaling are implicated in tumor growth. The hormone receptors expressed on breast cancer cells, including estrogen and progesterone receptors, allow complex crosstalk between hormonal pathways that drive cancer progression. The insights gained from this project could be used to develop helps to guide advance breast cancer prevention, diagnosis, and treatment decisions and predict responses to hormonal therapies (Satpathi et al., 2023)

2. Aim

The aim of this project is to investigate the relationship between hormonal imbalances, specifically estrogen and progesterone levels, and the incidence of breast tumors. Additionally, the project aims to determine the survival rates of individuals diagnosed with breast cancer in relation to their hormonal profiles.

3. Purpose

The purpose of this study is to examine a dataset of breast cancer patients to find the parameters linked with breast cancer survival rates. The dataset includes details regarding the patients' medical history, such as tumor size, hormone imbalances, tumor grade, and survival status. The study's goal is to create a predictive algorithm that can accurately forecast breast cancer survival rates based on medical data from patients. This study's findings could be used to create more effective treatment strategies and enhance patient outcomes.

Hormonal imbalance, notably estrogen and progesterone, has been extensively implicated in the development and progression of breast cancer (Satpathi et al., 2023). According to the study. Based on our findings, we believe that the status of hormone receptors (estrogen and progesterone receptors) influences breast cancer survival.

4. Background

Breast cancer is a global health challenge, affecting millions of women each year. While the impact of hormonal factors, like estrogen exposure, on breast cancer risk is well-established, their intricate relationship with tumor characteristics and patient outcomes remains unclear. Small studies have hinted at the role of hormones like estrogen, progesterone, and prolactin in breast cancer development and progression, but data limitations have hindered comprehensive analysis. Recent advances in hormonal assays and data analytics offer an exciting opportunity to unveil predictive hormonal patterns that could transform risk assessment, screening, prevention, treatment, and monitoring. To harness this potential, a large-scale analysis that integrates demographics, clinical records, tumor details, treatments, hormone levels, and survival outcomes is essential. The Kaggle breast cancer dataset provides a unique resource for this endeavor. By unraveling the complexities of hormonal influences, this project has the potential to drive innovations that alleviate the burden of breast cancer on patients, families, and society. (Feng et al., 2018)

5. Methodology:

5.1 Type of study:

This study is structured as an observational retrospective analysis, which means that no direct intervention or variable manipulation will be used in its analysis of historical data that has previously been gathered. It specifically uses a retrospective cohort study design, where the goal is to look back on the past to uncover factors linked to breast cancer survival, such as the medical histories of breast cancer patients. The dataset used in this study includes data on breast cancer patients' survival rates, which provides important insights into the elements that affect breast cancer survival. To find patterns and interactions between variables, rigorous observation and analysis of the available data are used in this research, which is observational in nature. (De Sanctis et al., 2022)

5.2 Data Collection:

Collected a comprehensive breast cancer dataset (about 4024 patients) that includes information on estrogen and progesterone levels, tumor characteristics, and patient survival.

Source: https://www.kaggle.com/datasets/alaahussien/breast-canser

5.3 Data Storage and Data Extraction:

All the data collected is in the CSV format, which will be stored securely in MySQL. Data extraction involves retrieving the relevant data from the breast cancer dataset, particularly estrogen and progesterone levels, along with other pertinent variables related to patient medical records using SQL queries.

5.4 Data Description:

Breast cancer data set containing 4024 rows and 16 columns including age, marital status, tumor grade, tumor size, and estrogen and progesterone, and live or death status. Our goal is to unveil statistically significant associations between hormonal imbalances and breast cancer outcomes. Normality test is used to determine if the data is following normal distribution or not.

5.5 Data Cleaning:

- At first, we start by identifying and addressing missing values in the dataset for variables such as estrogen and progesterone levels and filtering of the data with necessary columns related for the study by using SQL.
- In the breast cancer dataset, we will consider the following columns: age, tumor size estrogen levels, progesterone levels, survival months and status of live or death.
- In the breast cancer dataset, we will not consider the following columns: race, marital status, T, N, 6th, A stages, regional node examined, regional node positive, differentiate.
- Utilize the rename function to rename the columns in the breast cancer dataset and eliminate duplicate records to maintain data integrity and avoid duplicative information.
- Merge relevant tables or datasets and once the data preprocessing is completed, we will connect the final cleaned dataset to Python.

6. Data Analysis:

Conduct statistical analyses to examine associations between estrogen and progesterone levels and tumor incidence. Use survival analysis methodologies to determine how hormonal patterns influence patient survival over time.

6.1. Data Visualization:

- Utilizing data visualization tools and techniques such as scatter plots, histograms, and box plots will be used to visualize the relationships between the variables in the dataset and to identify any patterns or trends in the breast cancer data.
- Using libraries such as matplotlib and seaborn we intend to draw bar charts, scatter plot, ROC curve, learning and validation curves.

6.2. Machine learning:

- Machine learning models can be used to predict the likelihood of breast cancer occurrence based on risk factors and to identify the most important risk factors that contribute to breast cancer. This includes logistic regression, random forest, and support vector machine (SVM).

6.3. Statistical tests:

ANOVA, Regression, Normality test and correlation test.

- ANOVA can be used to compare the mean values of a risk factor across multiple groups, such as different age groups or hormonal imbalance and various stage of breast cancer.
- Regression analysis can be used to identify the relationship between a risk factor and the incidence of breast cancer.
- Normality tests can be used to check if the data follows a normal distribution, which is important for many statistical tests.
- Correlation tests can be used to identify the relationship between two risk factors and their impact on the incidence of breast cancer.

7. Hypothesis:

7.1 Scenario 1:

Null Hypothesis:

There is no significant association between estrogen and progesterone levels in breast cancer incidence.

Alternative Hypothesis:

There is a significant relation with hormonal imbalances specifically estrogen, and progesterone levels are associated with risk of breast cancer.

7.2 Scenario 2:

Null Hypothesis:

There is no significant association between hormonal induced breast cancer and survival rate.

Alternative Hypothesis:

There is a significant relation between hormonal induced breast cancer and survival rate.

8. Deliverables:

8.1 Data Collection and Preprocessing:

-Acquire raw breast cancer dataset from Kaggle containing patient information like demographics, hormonal levels, tumor characteristics, treatments, and survival outcomes.

-Store data securely in a SQL database for preprocessing.

-Clean data by handling missing values, duplicate records, and inconsistencies using SQL queries and Python scripts.

-Filter and select relevant columns like age, tumor size, estrogen/progesterone levels, survival status.

-Create cleaned, merged, and final dataset for analysis.

8.2 Exploratory Data Analysis:

-The Breast Cancer dataset from Kaggle contains information on 4024 patients with breast cancer, including variables such as age, tumor size, estrogen/progesterone levels, and survival status.

- Exploratory data analysis (EDA) was performed to understand the distribution of key variables and identify any correlations.

8.3 Summary Statistics:

Summary statistics will be generated for numerical variables like tumor size and hormone levels using Pandas and NumPy in Python. The mean tumor size was 22.53 mm (standard deviation = 13.49 mm), mean estrogen level was 135.58 pg/ml (standard deviation = 140.78 pg/ml), and mean progesterone level was 3.21 ng/ml (standard deviation = 4.91 ng/ml).

8.4 Statistical Analysis:

-Conduct ANOVA to compare hormone levels across tumor stages/grades.

-Perform regression analysis to identify relationships between hormones and tumor incidence.

-Apply correlation analysis to measure strength of associations.

-Use survival analysis methods like Kaplan-Meier curves.

-Test for normality and appropriate statistical assumptions.

8.5 Model Development:

-Split data into train, validation, and test sets.

-Develop logistic regression model to predict breast cancer incidence.

-Develop risk prediction model using regression or gradient boosting to estimate hormonal breast cancer risk.

-Tune model hyperparameters and cross-validate for optimal performance.

8.6. Visualizations:

-Visualizations will be created using Matplotlib and Seaborn to understand the distributions of key variables.

-Will show the distribution of tumor sizes, revealing a right-skewed distribution with a long tail. This indicates that most tumors are relatively small, but a small number of tumors are very large.

-Will illustrate the hormone levels through histogram overlays, showing estrogen with a right-skew and progesterone more normally distributed. This suggests that estrogen levels may be more variable than progesterone levels.

8.7. Correlations:

A correlation matrix heatmap will be generated to identify relationships between variables. Key findings include moderate positive correlations between tumor size and estrogen (r=0.42), as well as tumor grade and progesterone (r=0.31). This suggests that tumors with higher estrogen levels are more likely to be larger and have a higher grade.

8.8. Conclusion:

This initial EDA provides some descriptive insights into the Breast Cancer data. Key next steps will be conducting statistical hypothesis tests to further analyze the relationships between hormones and tumor characteristics. The visualizations and summary statistics will also inform the preprocessing and feature engineering steps for predictive modeling.

Results:

The project's results will provide insights into the relationship between hormonal imbalances (specifically estrogen and progesterone) and breast tumor incidence, as well as their impact on patient survival rates. It will also include predictive models that can help assess breast cancer risk and prognosis based on hormonal profiles.

Team members Responsibility:

Name of the team member	Background	Responsibilities
Priyankitha Kandi	Bachelor of Dental Surgery SQL Intermediate Level Python Intermediate Level	SQL and Python coding Data collection Data presentation Introducing initial views.
Kalyan Raj Chinigi	Doctor of Pharmacy SQL Intermediate Level Python Basic Level	Python coding Developing and implementing machine learning and cluster modules Provide required clinical insights.
Kavya Surabhi	Bachelor of Pharmacy SQL Intermediate Level Python Basic Level Data Visualization	SQL coding Data collection from data sets and assess data quality. Scheduling meetings and editing drafts.
Haneesha Chowdary Dondepudi	Bachelor of Pharmacy SQL Intermediate Level Python Basic Level	Python coding and presentation. Ensuring Data quality and uniqueness, data cleaning, handling discrepancies and pre-processing.
Srihari Myla Venkata	Doctor of Pharmacy SQL Intermediate Level Python Beginner Level	Python coding Data collection Data cleaning to eliminate null exclude null values by using python.

Saikumar Reddy Yalagalapalli	Master of Science in Food Sciences and Technology SQL Intermediate Level Python Basic Level Data Visualization	SQL coding Data visualization using python libraries. Writing drafts and submitting group assignments.

Timeline:

SL. No.	Date	Tasks
1	10/02 - 10/22	Data Examination and Cleansing • Identifying Null Values and Inconsistencies • Cleaning and Ensuring Data Consistency using Python. • Standardizing and Transforming Data for Analysis in Python
2	10/23- 11/15	Clustering Model Development and Validation • Using SQL, Python to clean the data, which includes handling missing values, outliers, and any data inconsistencies. • Performing data imputation or removal of rows/columns as necessary. • Developing Machine Learning Clustering Model • Train, Validate, and Test Model for Accuracy and Dependability
3	11/16- 11/20	Insight Visualization • Creating Data Visualizations with Seaborn and Matplotlib • Communicating Insights from Clustering
4	11/26- 11/30	Clustering and Risk score evaluation • Performing the clustering analysis. • Determining the optimal number of clusters and visualizing the results. • Assessing the risk scores generated by risk prediction model.
5	11/31- 12/04	Compiling findings, methodology, and results into a comprehensive final project report. • Detailed explanations of data preprocessing, modelling, and visualization processes. • Providing insights gained from the clustering analysis and risk assessment. • Considering recommendations or implications of the findings.

References:

(N.d.). Kaggle.com. Retrieved October 23, 2023, from https://www.kaggle.com/datasets/alaahussien/breast-canser.

De Sanctis, V., Soliman, A. T., Daar, S., Tzoulis, P., Fiscina, B., Kattamis, C., & International Network Of Clinicians For Endocrinopathies In Thalassemia And Adolescence Medicine Icet-A. (2022). Retrospective observational studies: Lights and shadows for medical

writers. *Acta Bio-Medica : Atenei Parmensis*, *93*(5), e2022319. https://doi.org/10.23750/abm.v93i5.13179

Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W., Liu, B., Lei, Y., Du, S., Vuppalapati, A., Luu, H. H., Haydon, R. C., He, T.-C., & Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & Diseases*, *5*(2), 77–106. https://doi.org/10.1016/j.gendis.2018.05.001

Pesheva, E. (2023, May 17). *Estrogen a more powerful breast cancer culprit than we realized*. Harvard Gazette. https://news.harvard.edu/gazette/story/2023/05/estrogen-a-more-powerful-breast-cancer-culprit-than-we-realized/

Satpathi, S., Gaurkar, S. S., Potdukhe, A., & Wanjari, M. B. (2023b). Unveiling the Role of Hormonal Imbalance in Breast Cancer Development: A Comprehensive Review. *Cureus*, *15*(7), e41737. https://doi.org/10.7759/cureus.41737