**Background**

Breast cancer is one of the most common malignancies in women worldwide and a leading cause of cancer-related mortality1. Despite advances in early detection and treatment, breast cancer remains a highly heterogeneous disease, with subtypes such as luminal A, luminal B, HER2-enriched, and triple-negative breast cancer (TNBC) showing variable prognosis and response to therapy2. While chemotherapy is a pillar of breast cancer treatment, its efficacy varies widely across tumor subtypes, stages, and patient populations. In particular, aggressive subtypes such as TNBC and advanced-stage breast cancer often have poor outcomes despite intensive chemotherapy regimens. In addition, chemotherapy can cause significant toxicity, impacting patients' quality of life and adherence3. The variability in response underscores the need for better predictive markers and tailored therapeutic strategies to optimize outcomes.

Survival analysis provides a robust framework for evaluating the impact of chemotherapy on long-term outcomes and identifying prognostic factors that can guide personalized treatment. By examining the relationship between survival, tumor characteristics, patient background and treatment regimens, this report aims to provide insight into the real-world efficacy of chemotherapy. Understanding these dynamics is critical to improving therapeutic strategies, particularly for patients with aggressive subtypes or poor prognosis, ultimately advancing efforts to reduce recurrence and improve survival in breast cancer patients.

**Methods**

***Exploratory Data Analysis***

To investigate the baseline characteristics and potential predictors of survival in breast cancer patients receiving chemotherapy, we first performed exploratory data analysis (EDA). The dataset included 1,977 patients stratified into groups with and without chemotherapy. Summary statistics for continuous variables, such as age and tumor size, were calculated and visualized using histograms and box plots (Figure 1) to identify central tendencies and distributions. Categorical variables, such as menopausal status and tumor stage, were summarized as proportions and analyzed for group differences using Pearson's chi-squared tests and Fisher's exact tests.

Correlations between continuous variables were assessed using pairwise Pearson correlation coefficients, as shown in a heat map (Figure 2). Statistical tests, including Wilcoxon rank-sum tests for non-normally distributed continuous variables, were used to assess differences between the chemotherapy and non-chemotherapy groups.

**Results**

***Exploratory Data Analysis***

EDA revealed significant differences in patient characteristics between the groups. Patients receiving chemotherapy were younger (mean age 50.5 vs. 63.9 years, p<0.001), more likely to be premenopausal (48% vs. 14%, p<0.001) and had larger tumors on average (mean size 32.3 mm vs. 24.7 mm, p<0.001). The distribution of tumor stages was skewed, with stage 2 and 3 being predominant among chemotherapy patients (74% and 19%, respectively).

Molecular subtype distributions also differed: basal-like, HER2-enriched, and claudin-low subtypes were overrepresented in the chemotherapy group, whereas luminal A dominated the non-chemotherapy group. Significant associations were observed between chemotherapy and key tumor characteristics, including ER status (negative: 62% vs. 14%, p<0.001) and histologic grade (grade 3: 79% vs. 42%, p<0.001).

**Discussion**

**Conclusion**

**References**

1. Trapani D, Ginsburg O, Fadelu T, et al. Global challenges and policy solutions in breast cancer control. *Cancer Treat Rev*. 2022;104(102339):102339.

2. Vranic S, Cyprian FS, Gatalica Z, Palazzo J. PD-L1 status in breast cancer: Current view and perspectives. *Semin Cancer Biol*. 2021;72:146-154.

3. Lake DE, Hudis CA. High-dose chemotherapy in breast cancer. *Drugs*. 2004;64(17):1851-1860.

Table 1. Baseline Characteristics of Patients with or without Chemotherapy

| **Characteristic** | **Without Chemotherapy**  N = 1,5651 | **With Chemotherapy**  N = 4121 | **P-value**2 |
| --- | --- | --- | --- |
| Age at diagnosis | 63.9 (12.0) | 50.5 (11.0) | <0.001 |
| Sex |  |  |  |
| Female | 1,565 (100) | 412 (100) |  |
| Menopausal status |  |  | <0.001 |
| Post | 1,339 (86) | 214 (52) |  |
| Pre | 226 (14) | 198 (48) |  |
| Cancer type |  |  |  |
| Breast | 15 (1.0) | 2 (0.5) |  |
| Breast Angiosarcoma | 0 (0) | 0 (0) |  |
| Breast Invasive Ductal Carcinoma | 1,169 (75) | 367 (89) |  |
| Breast Invasive Lobular Carcinoma | 129 (8.2) | 17 (4.1) |  |
| Breast Invasive Mixed Mucinous Carcinoma | 22 (1.4) | 1 (0.2) |  |
| Breast Mixed Ductal and Lobular Carcinoma | 193 (12) | 18 (4.4) |  |
| Invasive Breast Carcinoma | 36 (2.3) | 6 (1.5) |  |
| Metaplastic Breast Cancer | 1 (<0.1) | 1 (0.2) |  |
| Primary tumor laterality |  |  | 0.13 |
| Left | 783 (53) | 188 (49) |  |
| Right | 698 (47) | 199 (51) |  |
| Missing | 84 | 25 |  |
| Tumor size | 24.7 (12.6) | 32.3 (22.1) | <0.001 |
| Missing | 17 | 6 |  |
| Tumor stage |  |  | <0.001 |
| 0 | 11 (1.0) | 0 (0) |  |
| 1 | 480 (42) | 20 (6.3) |  |
| 2 | 591 (52) | 234 (74) |  |
| 3 | 58 (5.1) | 60 (19) |  |
| 4 | 7 (0.6) | 3 (0.9) |  |
| Missing | 418 | 95 |  |
| Cellularity |  |  | 0.004 |
| High | 739 (49) | 225 (56) |  |
| Low | 165 (11) | 50 (13) |  |
| Moderate | 611 (40) | 125 (31) |  |
| Missing | 50 | 12 |  |
| PAM50 claudin-low subtype |  |  |  |
| Basal | 96 (6.1) | 113 (27) |  |
| claudin-low | 142 (9.1) | 73 (18) |  |
| HER2 | 136 (8.7) | 88 (21) |  |
| LumA | 644 (41) | 56 (14) |  |
| LumB | 428 (27) | 47 (11) |  |
| NC | 6 (0.4) | 0 (0) |  |
| Normal | 113 (7.2) | 35 (8.5) |  |
| ER status |  |  | <0.001 |
| Negative | 218 (14) | 254 (62) |  |
| Positive | 1,347 (86) | 158 (38) |  |
| Histologic grade |  |  | <0.001 |
| 1 | 161 (11) | 8 (2.0) |  |
| 2 | 695 (47) | 76 (19) |  |
| 3 | 629 (42) | 323 (79) |  |
| Missing | 80 | 5 |  |
| HER2 status |  |  | <0.001 |
| Negative | 1,421 (91) | 309 (75) |  |
| Positive | 144 (9.2) | 103 (25) |  |
| Other histologic subtype |  |  |  |
| Ductal/NST | 1,136 (74) | 354 (87) |  |
| Lobular | 129 (8.4) | 17 (4.2) |  |
| Medullary | 13 (0.9) | 12 (3.0) |  |
| Metaplastic | 1 (<0.1) | 1 (0.2) |  |
| Mixed | 193 (13) | 18 (4.4) |  |
| Mucinous | 22 (1.4) | 1 (0.2) |  |
| Other | 15 (1.0) | 2 (0.5) |  |
| Tubular/ cribriform | 20 (1.3) | 1 (0.2) |  |
| Missing | 36 | 6 |  |
| Hormone therapy | 1,025 (65) | 191 (46) | <0.001 |
| Breast surgery |  |  | 0.032 |
| Breast Conserving | 641 (41) | 143 (35) |  |
| Mastectomy | 909 (59) | 260 (65) |  |
| Missing | 15 | 9 |  |
| Radiotherapy | 838 (54) | 335 (81) | <0.001 |
| Lymph nodes involvement | 1.4 (3.4) | 4.3 (5.5) | <0.001 |
| Missing | 58 | 16 |  |
| Mutations | 5.8 (4.2) | 5.2 (3.5) | <0.001 |
| Missing | 90 | 29 |  |
| Nottingham prognostic index | 3.8 (1.1) | 5.0 (0.9) | <0.001 |
| PR status |  |  | <0.001 |
| Negative | 635 (41) | 302 (73) |  |
| Positive | 930 (59) | 110 (27) |  |

1Mean (SD); n (%)

2Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test

Abbreviations: PAM50 (prediction analysis of microarray 50); ER (estrogen receptor); HER2 (human epidermal growth factor receptor 2); LumA (luminal A); LumB (luminal B); NC (not classified); NST (no special type); PR (progesterone receptor)

Figure 1. Histograms and Boxplots of Continuous Background Variables

A collage of multiple graphs

Description automatically generated

Figure 2. Correlations among Continuous Background Variables

A screenshot of a graph

Description automatically generated