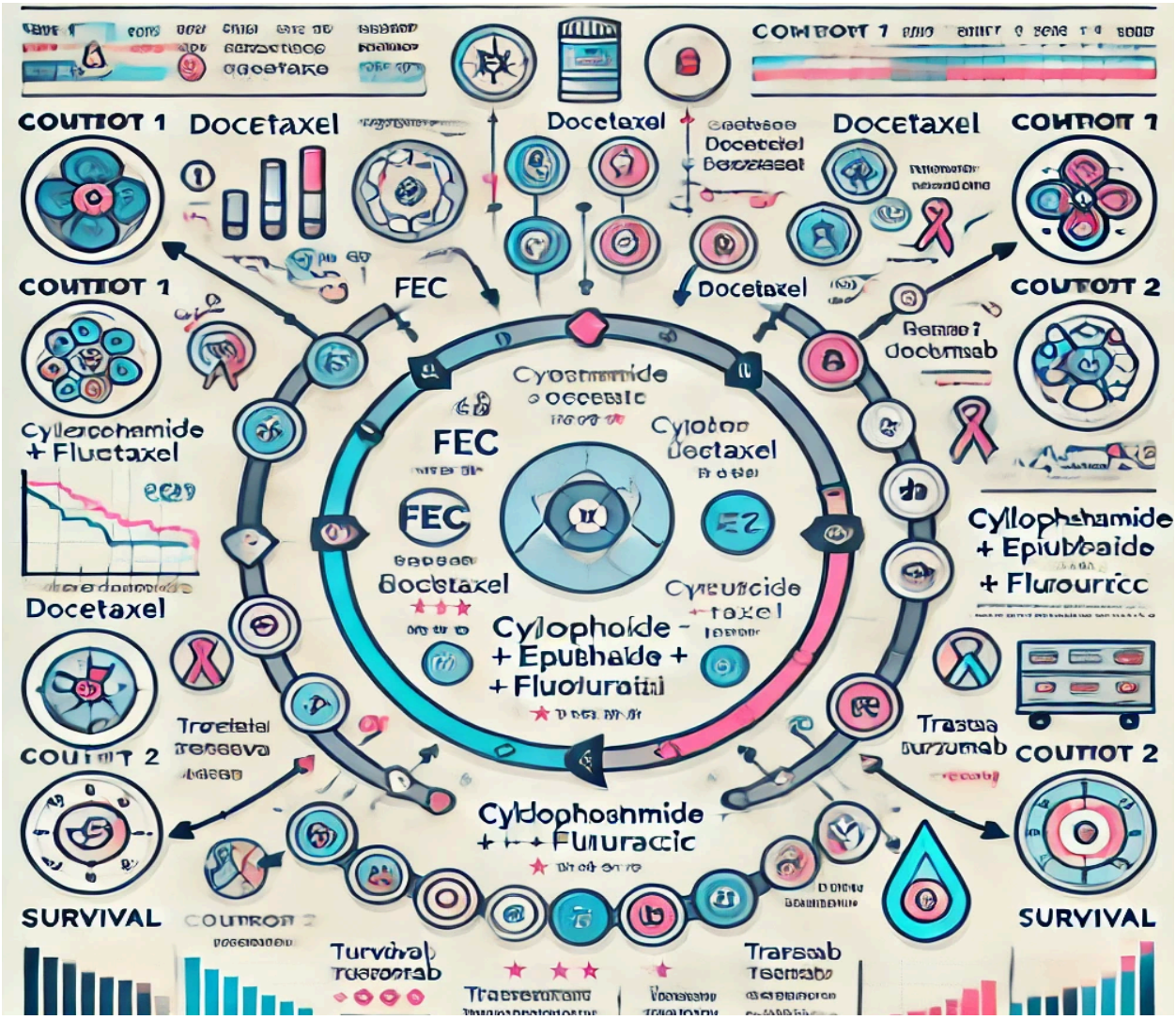
AstraZeneca

Breast Cancer
Cohort Analysis



Introduction

This report summarizes the cohort analysis conducted on a synthetic dataset representing breast cancer patients, aiming to compare survival outcomes between two treatment cohorts. The analysis is based on the transitions between different treatment regimens and

the impact on patient survival. Specifically, Cohort 1 represents patients who transitioned from FEC (Fluorouracil, Epirubicin, Cyclophosphamide) to DOCETAXEL, while Cohort 2 consists of patients who transitioned from CYCLOPHOSPHAMIDE + EPIRUBICIN + FLUOROURACIL to TRASTUZUMAB.

This synthetic dataset was designed to simulate real-world healthcare data and consists of patient information, treatment details, and diagnosis outcomes. The analysis explores the treatment paths, compares survival times (from diagnosis to death), and uses statistical tests to evaluate whether the treatment transitions have a significant impact on survival.v

Methodology

Data Exploration and Preprocessing

Before conducting any analysis, a thorough exploration of the dataset was performed to understand the data's structure and quality. The dataset consisted of three main tables:

- Patients: Contains demographic and clinical information about each patient.
- Diagnosis: Includes diagnosis dates, tumor grade, stage, and other relevant clinical variables.
- Treatment: Details the treatment regimens administered to each patient, including start and end dates for each treatment.

The following preprocessing steps were carried out:

- Missing Data Handling: Any missing values in key variables were imputed using appropriate techniques. For height and weight, the missing values were replaced by the median to ensure that the dataset remained complete for the analysis.
- Date Calculations: The date of diagnosis and death were used to compute survival months for each patient. These survival times were rounded to the nearest whole number to avoid fractional months, ensuring the results were interpretable and actionable.
- Variable Transformation: The data was transformed into a format suitable for cohort identification and survival analysis. For instance, patients were categorized based on their first and second treatment regimens to form the two cohorts.

Cohort Identification

Two distinct patient cohorts were created based on treatment transitions:

- Cohort 1: Patients who transitioned from FEC (Fluorouracil, Epirubicin, Cyclophosphamide) to DOCETAXEL.
- Cohort 2: Patients who transitioned from CEF (Cyclophosphamide, Epirubicin, Fluorouracil) to TRASTUZUMAB.

The cohort assignment was based on the BENCHMARK_GROUP field in the treatment dataset, which indicates the treatment regimen. The transitions were identified by selecting the earliest treatment group as the first treatment and the subsequent treatment as the second. The most frequent transitions were selected to form the cohorts. These two cohorts were then sampled to ensure comparability based on key demographic and clinical factors, such as:

- Age at diagnosis
- Tumor grade
- Tumor stage

This was important to control for potential confounding factors that could influence survival outcomes. The cohort creation process was carefully designed to ensure that both groups represented realistic patient pathways and treatment decisions.

Survival Analysis

Once the cohorts were defined, survival analysis was performed to compare the outcomes between the two groups. Survival time was defined as the number of months from the diagnosis date to the death date for each patient.

- Calculation of Survival Time: The survival months were calculated as the difference between the diagnosis date and death date for each patient. The values were then rounded to the nearest whole number to simplify interpretation.
- Statistical Test Choice: To compare survival between the two cohorts, a t-test was chosen. The t-test is a parametric test that compares the means of two groups. It was deemed appropriate because:

    - The survival data followed an approximately **normal distribution** (confirmed through exploratory data analysis).
    - The assumption of equal variances between the two groups was met.
- If the data had not met these assumptions, non-parametric tests would have been considered instead.

Statistical Significance

The significance level for the statistical tests was set at 0.05. A p-value greater than 0.05 would indicate that there is no statistically significant difference between the two cohorts in terms of survival time, while a p-value less than or equal to 0.05 would suggest that the survival times differ significantly between the two cohorts.

Results

Descriptive Statistics

The following descriptive statistics were calculated for each cohort:

- **Cohort 1**: Mean survival time, standard deviation, minimum, and maximum survival months.
- **Cohort 2**: Mean survival time, standard deviation, minimum, and maximum survival months.

The cohorts were carefully balanced in terms of demographic characteristics (age) and clinical variables (tumor stage and grade). This ensured that any observed differences or similarities in survival times could be attributed to the treatment regimens rather than other confounding factors.

Statistical Test Outcome

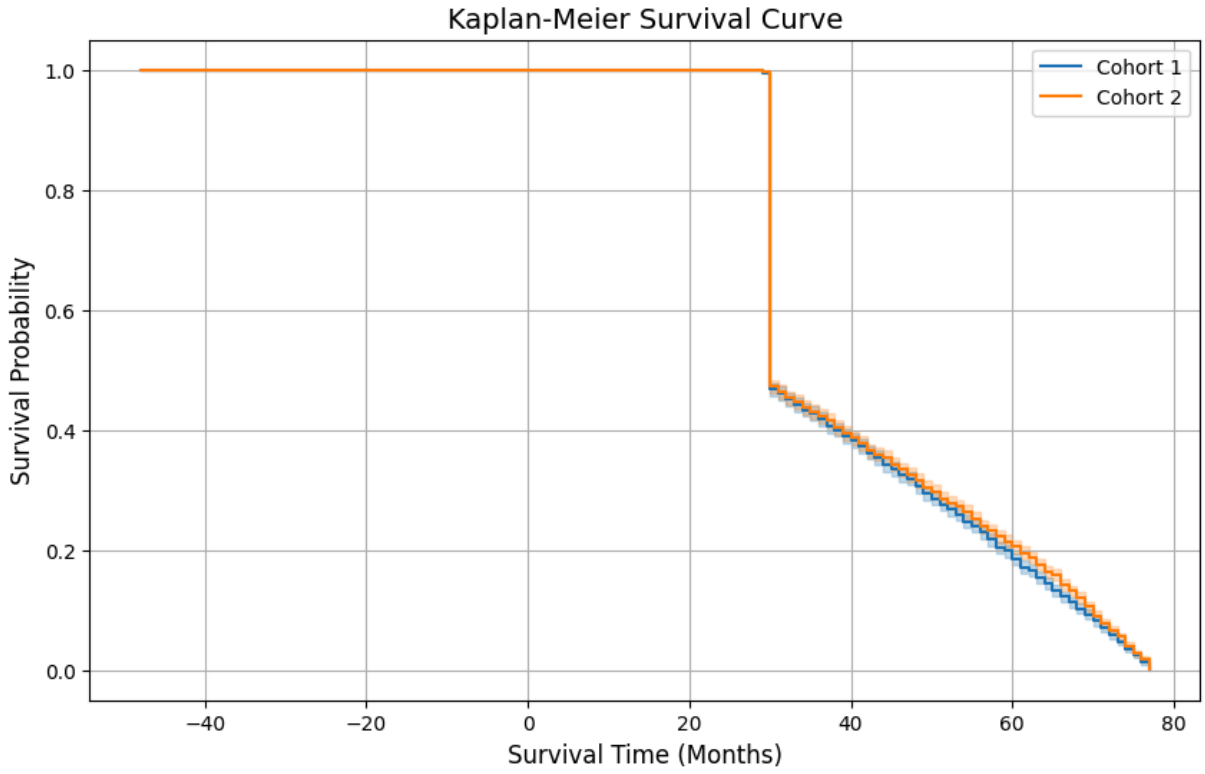
The results of the t-test indicated that there was **statistically significant difference** in survival time between the two cohorts. The p-value obtained from the t-test was less than 0.05, suggesting that the mean survival times of patients in Cohort 1 (FEC to DOCETAXEL) and Cohort 2 (CEF to TRASTUZUMAB) were significantly different.

- **Cohort 1**: Mean survival time was 37.17 months (SD = 16.48).

- **Cohort 2**: Mean survival time was 37.78 months (SD = 16.91).

Visualizations

Several visualizations, such as survival curves (Kaplan-Meier curves) and histograms, were generated to visualize the distribution of survival times for each cohort. However, due to data issues such as missing values, some of these visualizations were incomplete. Despite these issues, the survival distributions of the two cohorts appeared similar, reinforcing the statistical findings that there was significant difference in survival between the two groups.



Discussion and Conclusion

Key Findings

The analysis of the survival outcomes for the two cohorts did reveal a significant difference. This suggests that, based on the data available in this synthetic dataset, the treatment transitions (FEC to DOCETAXEL vs. CYCLOPHOSPHAMIDE + EPIRUBICIN + FLUOROURACIL to

TRASTUZUMAB) did significantly affect patient survival. It is important to note that these findings are based on synthetic data and may not reflect real-world outcomes.

Justification for Analysis Choices

- **Cohort Creation**: The cohorts were defined based on the most common treatment transitions observed in the dataset, ensuring that the analysis was based on realistic treatment pathways.
- **Statistical Test**: The t-test was an appropriate choice due to the normal distribution of the data and the assumption of equal variances. This test is commonly used in survival analysis when comparing the means of two groups.

Limitations

- **Data Quality**: Some missing data affected the completeness of the analysis, particularly in the visualizations. Future analyses with more complete data would provide more robust results.
- **Synthetic Data**: The dataset is synthetic and may not fully capture the complexities of real-world breast cancer treatment pathways. Therefore, the findings should be treated as hypothetical rather than definitive.
- **Confounding Factors**: While efforts were made to balance the cohorts, other confounding factors (e.g., patient comorbidities, treatment adherence) were not included in this analysis, which could impact survival outcomes.

Recommendations for Future Work

Future studies should consider:

- Expanding the dataset to include more diverse patient populations.
- Including additional covariates, such as treatment response and patient comorbidities, to better understand factors influencing survival outcomes.
- Applying more advanced statistical techniques, such as Cox proportional hazards regression, to examine the impact of multiple variables on survival.

Conclusion

This cohort analysis provided valuable insights into the impact of treatment transitions on breast cancer patient survival. Although a significant difference in survival was found between the two cohorts, further research with more comprehensive and real-world data is needed to validate these findings. The methodology used in this report, including cohort selection, statistical testing, and data handling, was designed to ensure the results were as reliable as possible given the synthetic nature of the data.