Technical Test

Methods

Data Acquisition and Merging

Patient clinical and survival data were obtained from multiple sources, including Body Composition Analysis (BCA) data and clinical records. The datasets were merged into a single structured dataset by matching unique patient identifiers.

Data Cleaning and Preprocessing

1. **Missing Data Reporting and Handling**
    a. A missing data report was generated to assess the extent of missing values in key variables.
    b. Multiple imputation was applied where necessary, using regression-based or mean/mode imputation strategies.
    c. Cases with excessive missingness in critical variables (e.g., survival time) were excluded from further analysis.
2. **Data Type Conversions**
    a. Survival time (Overall survival in months from diagnosis) was converted to a numeric format.
    b. Death during follow-up was encoded as a binary variable (1 = death, 0 = censored).
    c. Treatment groups were classified into adjuvant therapy (Yes/No) and neoadjuvant therapy (Yes/No) categories.

Exploratory Data Analysis (EDA)

Normality Testing for Body Composition Variables

To determine whether Total Muscle Volume (TMV) and Total Adipose Tissue Volume (TATV) followed a normal distribution, the Shapiro-Wilk test was conducted. Results guided subsequent statistical choices (parametric vs. non-parametric tests).

Summary Table for Clinical Variables

A summary table was generated to describe both continuous and categorical clinical variables:

- Continuous variables (e.g., age, BMI, muscle volume) were summarized using mean ± standard deviation (SD) or median (IQR) if skewed.

- Categorical variables (e.g., sex, therapy status) were presented as counts and percentages.

Survival Analysis

Kaplan-Meier Estimator

Kaplan-Meier survival curves were generated to compare survival probabilities between therapy groups:

- Separate survival functions were estimated for adjuvant therapy and neoadjuvant therapy groups.
- The log-rank test was used to compare survival distributions.

Cox Proportional Hazards Model

To quantify the impact of therapy and body composition on survival, Cox regression models were fitted:

1. Univariate analysis to assess each variable separately.
2. Multivariate models adjusting for covariates such as age, sex, BMI, and treatment type.
3. Hazard ratios (HR) and 95% confidence intervals were reported to evaluate risk differences.

Body Composition Analysis

Intramuscular Adipose Tissue (IMAT) / Total Muscle Volume (TMV) Ratio

The IMAT/TMV ratio was calculated as a potential predictor of survival outcomes. The clinical significance of this ratio was assessed using:

- Receiver Operating Characteristic (ROC) curve analysis to determine an optimal cutoff for survival prediction.
- Kaplan-Meier survival curves stratified by IMAT/TMV threshold.

Impact of Body Composition on Overall Survival

To assess how muscle mass and adiposity influence survival, Cox regression models were used to analyze:

- Total Muscle Volume (TMV)
- Total Adipose Tissue Volume (TATV)
- IMAT/TMV Ratio

Visualization and Statistical Comparisons

Visualization of Gender & Age-Related Differences

To explore the effect of gender and age on body composition, the following visualizations were created:

- Boxplots and violin plots comparing TMV, TATV, and IMAT/TMV ratios by sex and age group.
- Scatter plots examining correlations between age and body composition parameters.

Kaplan-Meier and Cox Model Visualizations

- Kaplan-Meier survival plots were used to display differences in survival curves.
- Forest plots visualized hazard ratios from Cox regression models.

Statistical Software & Significance Thresholds

All analyses were conducted using Python, with pandas for data processing, lifelines for survival analysis, and seaborn/matplotlib for visualization. Statistical significance was set at $p < 0.05$ for all tests.