Project Proposal A : Predicting Long-Term Recurrence Risk in Breast Cancer Using a Multimodal Model Integrating Clinical and Genomic Data

Baodong Zhang 20.06.2025

Objective

To build a multivariable survival model that integrates clinical features and selected gene expression profiles to predict long-term recurrence risk in breast cancer patients.

Three models will be developed and compared: A clinical-only model; A gene expression-only model; A combined multimodal model. Performance will be evaluated to explore the potential improvement when combining clinical and genomic features.

Data Source

Dataset: GSE7390 from NCBI GEO

Cohort: 198 lymph node-negative breast cancer patients

Clinical variables: age, tumor grade, ER (estrogen receptor) status, tumor size etc.

Gene expression: microarray data with thousands of probes

Outcome: recurrence status and time to recurrence (or overall survival time)

Data Preprocessing and Exploratory Data Analysis (EDA)

Clean clinical data: handle missing values and outliers

Normalize gene expression data: e.g., log2 transformation, or z-score standardization.

Perform unsupervised clustering (e.g., K-means or hierarchical clustering) on gene expression data to identify potential molecular subgroups. Use t-SNE or UMAP for 2D visualization of clustering patterns to aid interpretation of subgroup structure.

Feature Selection (if proportional hazards (PH) assumption holds)

Method 1: Univariate Cox analysis to identify survival-associated genes

Method 2: LASSO-Cox regression to select key predictive genes

Model Development

Three Cox proportional hazards models will be constructed:

Clinical model: age + tumor grade + ER status + tumor size + etc.

Genomic model: selected gene expression features (e.g., top 30–100 genes)

Combined model: clinical variables + selected gene features

The risk score for patient will be calculated as:

$$\text{Risk Score}_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Patients will be stratified into high- and low-risk groups (e.g., using median split).

Model Evaluation

C-index: concordance between predicted risk and actual survival time ranking

Time-dependent ROC/AUC: model sensitivity and specificity at various time points

Kaplan-Meier survival curves: visualize survival differences between risk groups

Use log-rank test to assess significance

External validation may be performed using an independent dataset with similar structure. If similar dataset is not available, train test split will be performed.

Challenges

High-dimensional genomic data: Requires careful feature selection to avoid overfitting, noise, and computational burden

PH assumption violation: May limit the use of Cox models;

Further Exploration (options)

Molecular subgroups identified through unsupervised clustering in the EDA phase will be evaluated for association with survival outcomes using Kaplan–Meier curves and the log-rank test. If significant, cluster labels or principal components will be included as covariates in multivariate survival models to assess their prognostic value.

If the Proportional Hazards Assumption Is Violated, apply dimensionality reduction (e.g., PCA) and consider alternative models such as Random Survival Forests (RSF).

Project Proposal B: Machine Learning for Predicting Gallstone Disease Using Non-Imaging Clinical Data

Baodong Zhang 20.06.2025

Objective

The aim is to develop and evaluate machine learning models that predict the presence of gallstone disease using non-imaging clinical features. This high-sensitivity diagnostic tool will support early detection and reduce missed cases in clinical settings.

Data Source

Dataset: UCI Machine Learning Repository – Gallstone Disease Dataset

Sample Size: 319 individuals Positive cases: 161 diagnosed with gallstone disease

Features (total: 38): Demographics: age, gender, etc. Body composition: bioimpedance measurements. Laboratory values: metabolic and biochemical markers

Data Preprocessing and Analysis (EDA)

Data Cleaning: handle missing values, outliers, and inconsistent entries

Feature Selection: LASSO regression or decision trees, and statistical tests

Multivariate analysis: explore interactions between metabolic features

Dimensionality Reduction: apply PCA to body composition features

Confounding Analysis: assess gender-related bias and other confounders

Train-Test Split: 70% vs 30%

Model Development

Baseline Model – Logistic Regression. Serve as a reference model to establish baseline performance for comparison with more complex approaches.

Random Forest. Apply cross-validation for hyperparameter optimization and assess feature importance.

Hierarchical Bayesian Model. Estimate patient-specific posterior probabilities, account for gender-related bias in prediction, and provide uncertainty quantification to support clinical decision-making.

Model Evaluation

Sensitivity (Recall): Prioritized to minimize missed diagnoses and ensure early detection.

Specificity & Accuracy: Evaluate overall correctness and ability to distinguish negative cases.

AUC-ROC: Measures the model's discrimination ability across different classification thresholds.

Challenges

Handle continuous variables through binning or feature pre-selection to prevent excessive split points in tree models.

Choosing appropriate priors in Bayesian models is difficult without strong domain knowledge and can affect model reliability.

Bayesian models can be resource-intensive and require careful tuning for convergence.

Small dataset (n=319) increases the risk of overfitting, especially for complex models.

Balancing model performance with clinical interpretability is essential for real-world use.