Progress Report on Project Work

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

Baodong Zhang 25.07.2025

As of today, the following work has been completed:

1. Data Source and Preprocessing

This study is based on the GSE7390 dataset (n = 198) for model development and GSE2990 (n = 155) as the external validation cohort. The data includes:
- Clinical data: time to distant metastasis and event (time_dmfs, event_dmfs), age, estrogen receptor (ER) status, tumor grade, tumor size, and lymph node status
- Gene expression data: Affymetrix microarray platform

Clinical data preprocessing steps:
- Removed samples with excessive or critical missing values
- converted categorical variables to factors
- Applied MICE (Multiple Imputation by Chained Equations):
  - Ordered categorical variables: proportional odds logistic regression
  - Binary variables: logistic regression
  - Five imputations performed; mode used as final value
- No significant outliers detected via boxplots

2. Clinical Variable Modeling and Survival Analysis

- Selected clinical predictors: grade, size, age, ER status
- A Cox proportional hazards model was fitted initially; however, the global test from cox.zph indicated that the proportional hazards assumption was violated (p = 0.00118)
- Therefore, a Random Survival Forest (RSF) was used instead:
  - Cross-validation (CV) applied for parameter tuning
  - Model performance on training data:
    - iAUC = 0.8183
    - C-index = 0.7136
  - Performance on external dataset (GSE2990):
    - iAUC = 0.6554
    - C-index = 0.6166

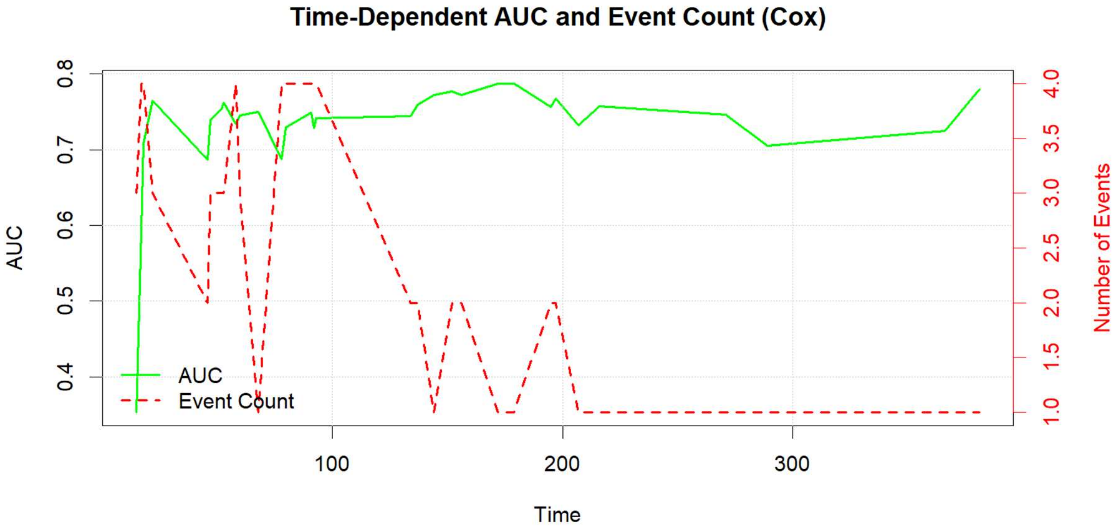
3. Gene Expression + Clinical Variable Modeling

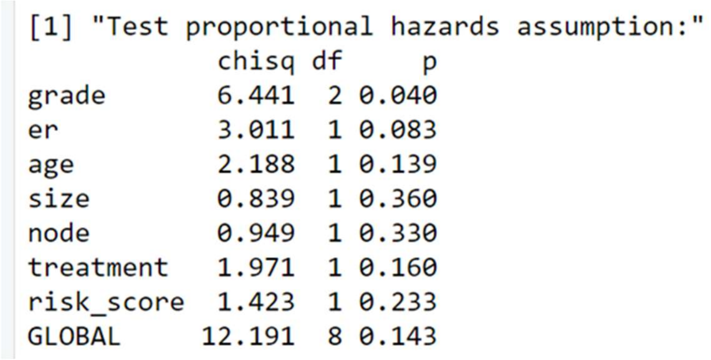
- The expression values in GSE7390 ranged from -2.85 to 16.91, indicating likely log2 transformation
- All genes were standardized using scale() (subtracting the mean and dividing by standard deviation)
- For validation GSE2990 was normalized using the same mean and standard deviation from GSE7390

Feature selection and risk score development:
1. Univariate Cox regression was performed on all 22,283 genes. About 3,068 genes had p-values ≤ 0.11
2. LASSO-Cox modeling based on the selected genes:
   - Optimal lambda = 0.09; 39 genes with non-zero coefficients were selected
   - Refitted Cox model using these genes; 15 genes with p < 0.05 were retained
   - A risk score was constructed using the coefficients of these 15 genes

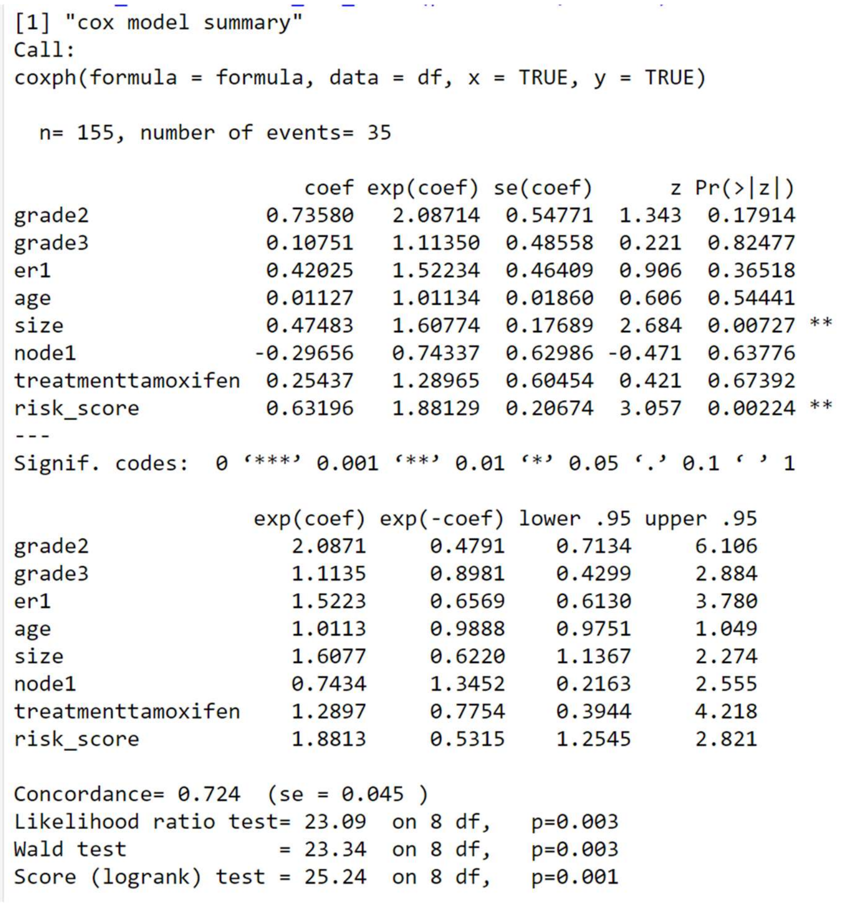
$$\text{Risk Score}_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

External validation (GSE2990):
- After applying the clinical features and risk score from GSE7390 on cox model, cox.zph indicated that the proportional hazards assumption was not violated (p = 0.143):
  - iAUC improved by 5.8% (from 0.6792 to 0.7372)
  - C-index improved by 3.5% (from 0.6974 to 0.7328)





```
[1] "Test proportional hazards assumption:"
             chisq df     p
grade        6.441  2 0.040
er           3.011  1 0.083
age          2.188  1 0.139
size         0.839  1 0.360
node         0.949  1 0.330
treatment    1.971  1 0.160
risk_score   1.423  1 0.233
GLOBAL      12.191  8 0.143
```



```
[1] "cox model summary"
Call:
coxph(formula = formula, data = df, x = TRUE, y = TRUE)

  n= 155, number of events= 35

                        coef exp(coef) se(coef)      z Pr(>|z|)
grade2               0.73580   2.08714  0.54771  1.343  0.17914
grade3               0.10751   1.11350  0.48558  0.221  0.82477
er1                  0.42025   1.52234  0.46409  0.906  0.36518
age                  0.01127   1.01134  0.01860  0.606  0.54441
size                 0.47483   1.60774  0.17689  2.684  0.00727 **
node1               -0.29656   0.74337  0.62986 -0.471  0.63776
treatmenttamoxifen   0.25437   1.28965  0.60454  0.421  0.67392
risk_score           0.63196   1.88129  0.20674  3.057  0.00224 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
grade2                2.0871     0.4791    0.7134     6.106
grade3                1.1135     0.8981    0.4299     2.884
er1                   1.5223     0.6569    0.6130     3.780
age                   1.0113     0.9888    0.9751     1.049
size                  1.6077     0.6220    1.1367     2.274
node1                 0.7434     1.3452    0.2163     2.555
treatmenttamoxifen    1.2897     0.7754    0.3944     4.218
risk_score            1.8813     0.5315    1.2545     2.821

Concordance= 0.724  (se = 0.045 )
Likelihood ratio test= 23.09  on 8 df,   p=0.003
Wald test            = 23.34  on 8 df,   p=0.003
Score (logrank) test = 25.24  on 8 df,   p=0.001
```

4. Exploratory Gene Analysis and Dimensionality Reduction

- PCA was applied to reduce the dimensionality of gene expression data
- A Variational Autoencoder (VAE) was trained with cross-validation to select the optimal latent dimension; however, it did not improve survival prediction
- UMAP and t-SNE were also applied for visualization, but no strong structural patterns were found that enhanced predictive performance

5. Current Issues and Questions

1. Is the gene expression preprocessing appropriate?
   - The data appears to be log2-transformed. Is it appropriate to apply additional standardization (scale), or are there better alternatives?

2. Model validation strategy (CV required by project):
   - A: Full cross-validation from the beginning — each fold includes a train/test split, and all modeling steps (e.g., LASSO CV) are done strictly within the training set to avoid data leakage
   - B: Build the model on the entire dataset, use internal CV to tune parameters, and validate performance only on the external test set
   - → Which strategy is more appropriate, especially given the small sample size (n = 198)?

3. Are there any methods currently used that may be inappropriate or could be improved?

4. Are there other feature selection / modeling techniques for survival analysis worth exploring to improve performance?

5. Are there additional directions to explore for improving performance using VAE, UMAP, or t-SNE?

6. Target variable selection:
   - Currently using time_DMFS and event_DMFS (distant metastasis-free survival) as the outcome. The dataset also includes time_RFS and event_RFS (recurrence-free survival).
   - Is one of these endpoints preferred? Or should both be tested and compared?