Project Progress Summary 2

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

Baodong zhang 31.07.2025

1. Missing Data and Imputation

GSE7390 had very few missing values, while GSE2990 had more. I applied MICE imputation to fill missing data in GSE2990.

2. Batch Effect Removal and Validation Issue

I used the Combat() function from the sva package to remove batch effects between GSE7390 and GSE2990, followed by scaling. However, external validation performance was poor (time AUC and C-index mostly between 0.50 and 0.60). Since the survival outcome (time/event) cannot be included in Combat(), important survival-related signals might have been removed as batch effects. Therefore, I abandoned external validation and switched to a full internal split (0.7 training / 0.3 test) using only GSE7390. The training data now includes ~70% of 198 samples.

3. Univariate Cox Selection Threshold

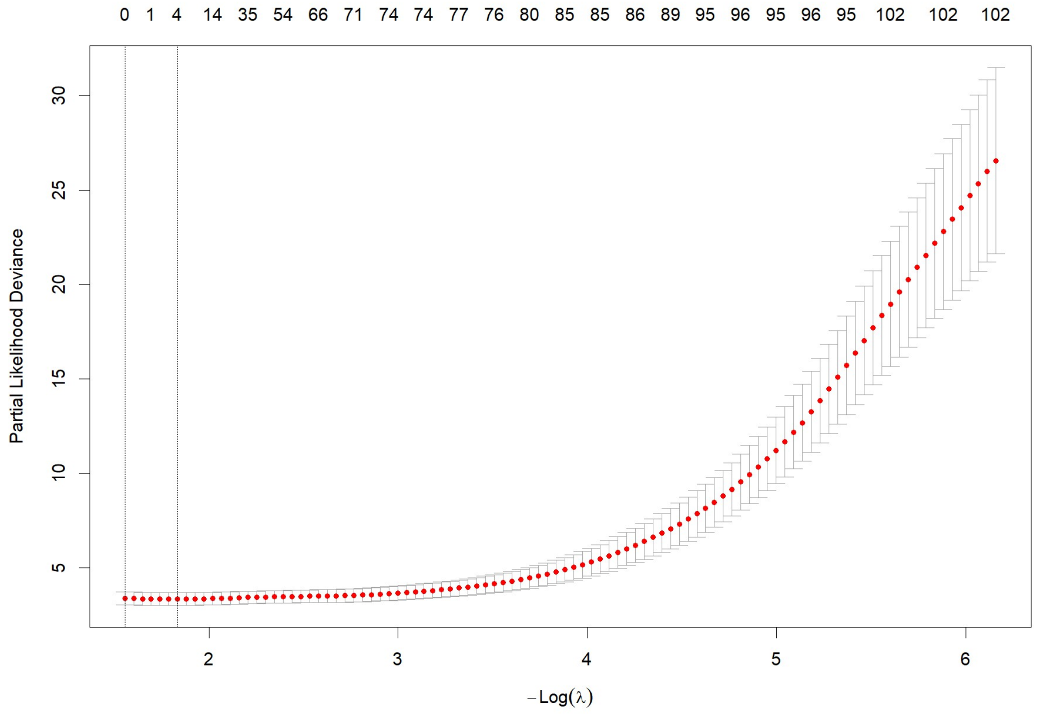
I originally used a p-value threshold of 0.11 in univariate Cox regression to include more cancer-related genes (e.g., HER2, EGFR). This was later changed to a more standard cutoff of 0.05.

4. Post-LASSO Cox Fitting

I initially added an extra Cox fitting step after LASSO to reduce the number of predictors. This sometimes improved validation performance, but I later realized it was likely due to data instability caused by the small sample size. This step has been removed.

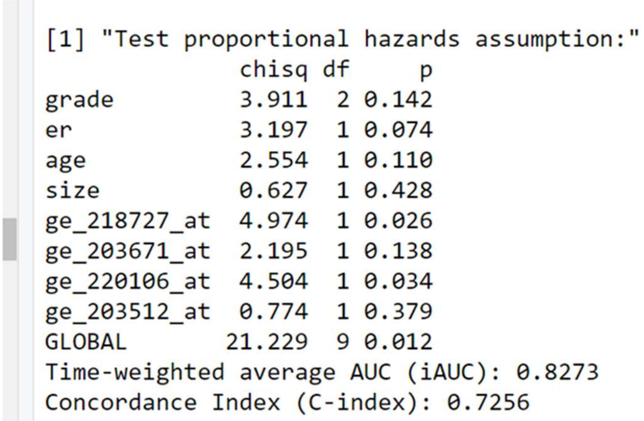
5. Current Models (Preliminary Results)

Two models have been established based on repeated experiments:
- Model 1: LASSO-selected genes + clinical variables → Cox model

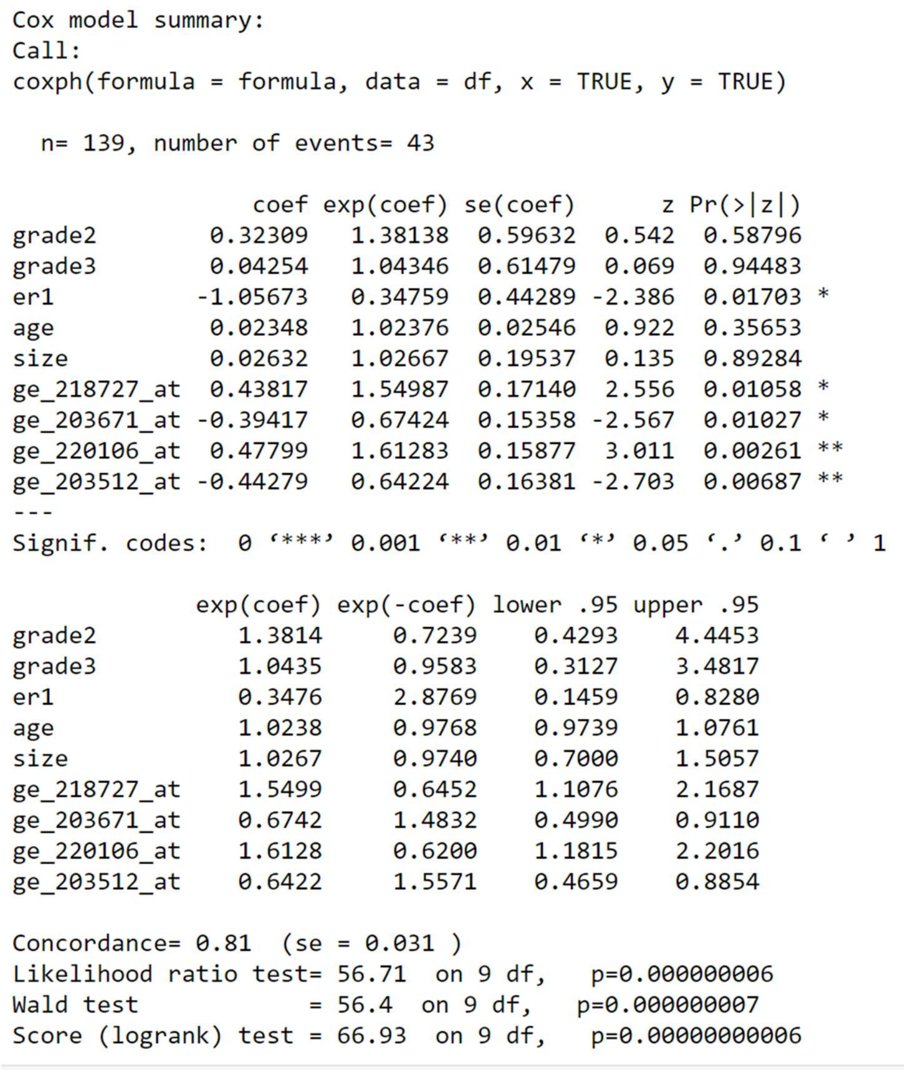


In the LASSO-Cox plot above, we can see that as the penalty becomes smaller, more genes are selected, but the partial likelihood deviance actually increases. This may be because the sample size is small, and using too many variables causes overfitting on the validation sets during cross-validation. In the end, only four genes were selected.

Directly using these key genes as individual variables in the model performed better than using a combined risk score calculated from multiple genes and their coefficients. On the test dataset (n ≈ 59), the model achieved an iAUC of 0.8273 and a C-index of 0.7256. However, the global test from cox.zph showed violation of proportional hazards assumption (p = 0.012)

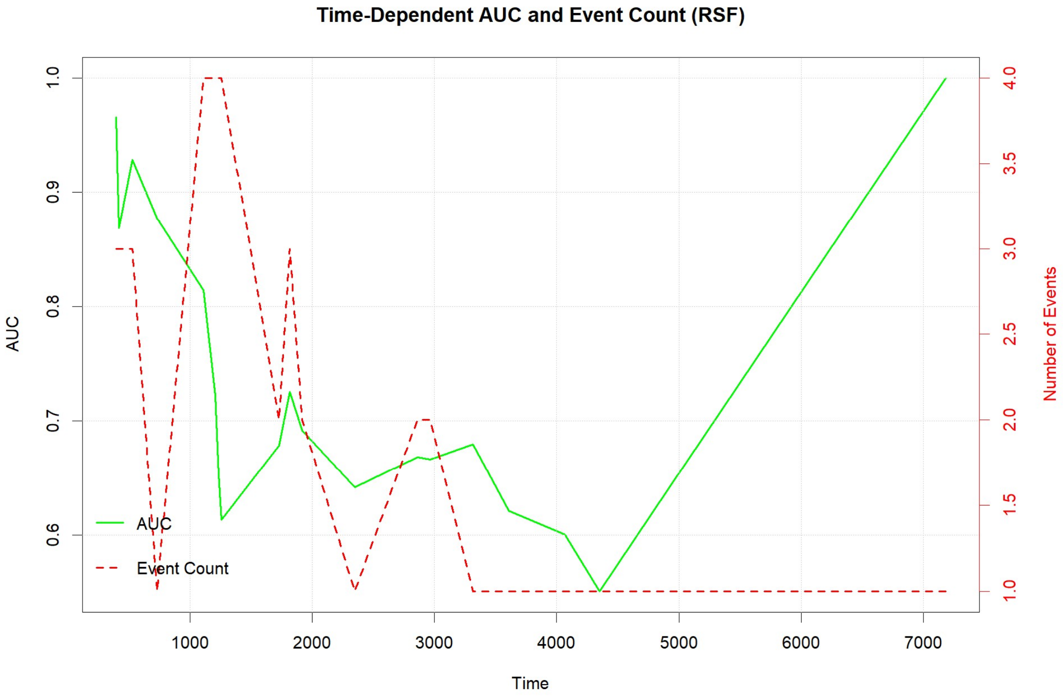


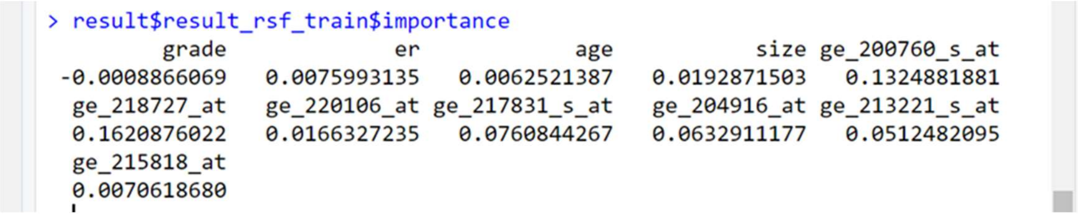
```
[1] "Test proportional hazards assumption:"
              chisq df     p
grade         3.911  2 0.142
er            3.197  1 0.074
age           2.554  1 0.110
size          0.627  1 0.428
ge_218727_at  4.974  1 0.026
ge_203671_at  2.195  1 0.138
ge_220106_at  4.504  1 0.034
ge_203512_at  0.774  1 0.379
GLOBAL       21.229  9 0.012
Time-weighted average AUC (iAUC): 0.8273
Concordance Index (C-index): 0.7256
```



```
Cox model summary:
Call:
coxph(formula = formula, data = df, x = TRUE, y = TRUE)

  n= 139, number of events= 43 

                 coef exp(coef) se(coef)      z Pr(>|z|)   
grade2        0.32309   1.38138  0.59632  0.542  0.58796   
grade3        0.04254   1.04346  0.61479  0.069  0.94483   
er1          -1.05673   0.34759  0.44289 -2.386  0.01703 * 
age           0.02348   1.02376  0.02546  0.922  0.35653   
size          0.02632   1.02667  0.19537  0.135  0.89284   
ge_218727_at  0.43817   1.54987  0.17140  2.556  0.01058 * 
ge_203671_at -0.39417   0.67424  0.15358 -2.567  0.01027 * 
ge_220106_at  0.47799   1.61283  0.15877  3.011  0.00261 **
ge_203512_at -0.44279   0.64224  0.16381 -2.703  0.00687 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
grade2          1.3814     0.7239    0.4293    4.4453
grade3          1.0435     0.9583    0.3127    3.4817
er1             0.3476     2.8769    0.1459    0.8280
age             1.0238     0.9768    0.9739    1.0761
size            1.0267     0.9740    0.7000    1.5057
ge_218727_at    1.5499     0.6452    1.1076    2.1687
ge_203671_at    0.6742     1.4832    0.4990    0.9110
ge_220106_at    1.6128     0.6200    1.1815    2.2016
ge_203512_at    0.6422     1.5571    0.4659    0.8854

Concordance= 0.81  (se = 0.031 )
Likelihood ratio test= 56.71  on 9 df,   p=0.000000006
Wald test            = 56.4  on 9 df,   p=0.000000007
Score (logrank) test = 66.93  on 9 df,   p=0.00000000006
```

- Model 2: LASSO-selected genes + clinical variables → Random Survival Forest (RSF)





```
> result$result_rsf_train$importance
        grade            er           age          size ge_200760_s_at
-0.0008866069  0.0075993135  0.0062521387  0.0192871503   0.1324881881
  ge_218727_at  ge_220106_at ge_217831_s_at  ge_204916_at ge_213221_s_at
  0.1620876022  0.0166327235   0.0760844267  0.0632911177   0.0512482095
  ge_215818_at
  0.0070618680
```

An RSF model composed of four genes and multiple clinical variables yielded similar results on the validation set - iAUC = 0.8160, C-index = 0.7151. Based on the variable importance from the above plot, it can be anticipated that removing "grade" will improve the model's performance.

6. Final Model: Selection, Evaluation, Application, and Interpretation

My planned final model is a Cox proportional hazards model that includes several clinical variables and a few genes. The model determines the coefficients (β) for each variable. By estimating the baseline hazard (Hazard_0), it can predict a patient's survival probability at different time points. The model has been evaluated on the test set by comparing predicted survival probabilities at different time points with actual outcomes (to compute time-dependent AUC and C-index).

The Cox model is easy to interpret because the size of each β coefficient indicates the contribution of the corresponding variable to the risk.

Question: If the Cox model violates the proportional hazards assumption ($p = 0.012$), should it be discarded entirely?

If so, I will try training the model on other datasets, such as GSE2990, to see whether the assumption still fails.

The RSF (Random Survival Forest) model may also be considered as the final model, although its interpretation is more complex.

For model application: A new sample needs to provide ER status, age, tumor size, and expression values of specific genes (e.g., ge_200760_s_at, ge_218727_at, etc.). The model can then predict survival probabilities over time.

7. Exploratory Gene Analysis and Dimensionality Reduction

I explored the use of dimensionality reduction (e.g., latent variable extraction) to replace gene selection. The idea is to use latent variables from PCA, VAE, or similar techniques as inputs for Cox or RSF models. However, initial experiments failed to identify meaningful latent features.

8. Next Steps

I plan to apply LASSO-Cox with bootstrap resampling to obtain more stable gene selection. I will also continue exploring other strategies to improve model performance.