Name: Yikang(Kevin) Zhang Student number: 29570447 BIOF520 Assignment2:
Github link: https://github.com/KevinZhang20010302/BIOF520_assignment2.git

Abstract:

The aim of this assignment is to develop a classification model to improve the clinical stratification of recurrence free survival in low grade Ta bladder cancer patient. And the performance should be better than the current EAU risk stratification for recurrence risk with meaningful clinical impact.

Two datasets were used in this study, UROMOL cohort dataset and Knowles cohort. Due to the gene expression data in UROMOL cohort dataset is RNA-seq data, in Knowle cohort dataset is microarray data, normalization to eliminate the batch effect is required[1]. In addition, clinical information differed between UROMOL and Knowles and ground truth of classification labels were missing. Therefore, we start with data preprocessing. Consider clinical impact, a penalized Cox model and a cox proportional hazards model were combined as the prediction model. Final classification was based on the predicted probability of recurrence on month 24. Model evaluation including both internal evaluation and external evaluation. The metric we choose for evaluation is Kaplan-Meier analysis, C-index for first penalized cox model and AUC for final classification model. We also calculate AUC for EAU classification to compare with our model.

Data preprocessing:

The data preprocessing including two parts: data preprocessing for clinical data and data preprocessing for gene expression data. First, we start with data preprocessing for gene expression data. There are two problems to solve: batch effect and differences in gene coverage between datasets. First, only genes present in both datasets were kept. Next, batch effect correction was performed by removeBatchEffect() from limma (version 3.66.0)[2].
Next step is the data preprocessing for clinical data. There are several issues exist in clinical data: first is missing value. Second is difference coverage of clinical information between two datasets. Third is non-numeric variables and survival data. Forth is irrelevant clinical variables. Fifth is missing ground truth classification labels. To address those problems. First, for missing data on survival data (RFS time + recurrence), if both RFS time and recurrence is missing, the sample was removed. If recurrence is not missing but RFS time is missing, the missing RFS data was imputed using the following up time (FUtime_days) divided by 30 to convert days to months. For other clinical variables, samples with missing values were removed.
Next, only clinical information present in both datasets were kept. Non-numeric variables were first converted to factors and then transformed as binary variables using matrix.model(). A special correction for values of UROMOL2021.classification in Knowles dataset as it is not consistent with UROMOL2021.classification in UROMOL dataset. RFS time and recurrence were convert as survival data by Surv() in rms version 8.1-0[3]. Ground truth classification labels were based on RFS time and recurrence: if RFS time > 24, mark as low risk, RFS <= 24 with recurrence 0, mark as unknown. RFS <=24 with recurrence 1, mark as high risk.
Finally, the preprocessed clinical data were combined with preprocessed gene expression data.

Model building:

After data preprocessing, the UROMOL dataset was spilt into train set (80%) and test set (20%). The penalized cox model (cox_penalty) was built by glmnet() in glmnet version 4.1-10 with

family as cox, alpha as 1 (lasso) and lambda.min selected using cv.glmnet() to predict risk score based on clinical and gene information of UROMOL train datasets[4,5].The cox proportional hazard model (no_recurrence_model) was built by coxph() in survival version 3.8-6 based on survival train data and risk score of train set[6].When using the model on new data, the risk score of new data were predicted by predict() with cox_penalty and the no recurrence probability across time were predicted using survfit() in survival version 3.8-6 with no_recurrence_model[6]. For classification, the tentative threshold of recurrence probability 0.5 at 24 months was used, higher probability means low risk.

Internal evaluation:

Kaplan-Meier plot: first KM plot below

AUC is 0.59. (roc() from pROC version (1.19.0.1))[7]

C-index is 0.56 (concordance() from survival version 3.8-6)[6].

External evaluation:

Kaplan-Meier plot: second KM plot below

AUC is 0.58.

C-index is 0.54

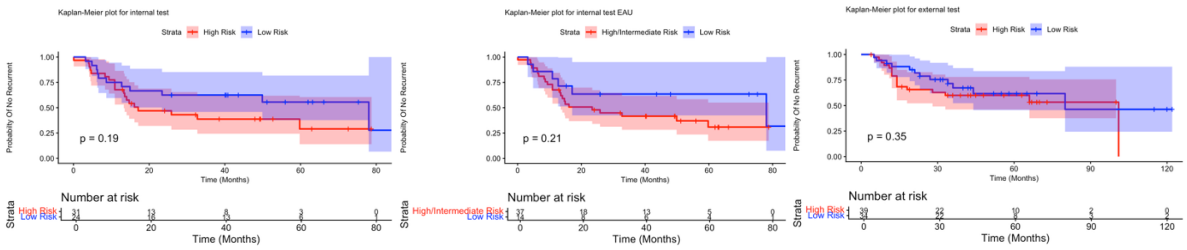
Compare with EAU:

Kaplan-Meier plot: third KM plot below

AUC is 0.56.

Clinical interpretation:

This model provides risk classification and shows probability of recurrence at defined time (month number). With the output, clinicians can determine whether the current treatment is effect or not, furthermore, help to make adjustment on the treatment.



Reference

1. Castillo, D. *et al.* Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling. *BMC Bioinformatics* **18**, 506 (2017).

2. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

3. Lenth, R. V. Response-Surface Methods in R, Using rsm. *J. Stat. Softw.* **32**, 1–17 (2010).

4. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

5. Tay, J. K., Narasimhan, B. & Hastie, T. Elastic Net Regularization Paths for All Generalized Linear Models. *J. Stat. Softw.* **106**, 1–31 (2023).

6. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer, New York, NY, 2000). doi:10.1007/978-1-4757-3294-8.

7. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).