



Development and validation of a multi-omics model for predicting response and survival of NSCLC Patients receiving Third-generation EGFR-TKI

Sitong Li¹, Ning Lou¹, Xiaohong Han^{1*}¹.Clinical Pharmacology Research Center, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing/CN

Background

Third-generation Epidermal growth factor receptor - tyrosine kinase inhibitors (EGFR-TKI) are crucial for the treatment of EGFR-mutant non-small cell lung cancer (NSCLC). However, due to patient heterogeneity, not all NSCLC patients with EGFR-sensitizing mutations respond to EGFR-TKI therapy, with approximately 30% of patients failing to respond to third-generation EGFR-TKI. The majority of patients will inevitably develop acquired resistance, leading to disease progression. Currently non-invasive biomarkers are lacking for efficacy and prognosis prediction. This study aims to integrate proteomics and metabolomics, revealing the potential non-invasive biomarkers for response to third-generation EGFR-TKI in NSCLC patients.

Method

We conducted pre-treatment plasma metabolomic and proteomics analysis on 161 NSCLC patients in two independent cohorts. Cohort 1 containing 131 patients was randomly split 7:3 into training (n=93) and internal validation (n=38) sets. Cohort 2 was assigned as test set (n=30). Feature selection was performed via multi-omics feature integration and Spearman correlation analysis, and 4 machine learning models were trained. We evaluated the predictive performance for therapeutic response using AUC, and assesses survival outcomes including overall survival (OS) and progression-free survival (PFS).

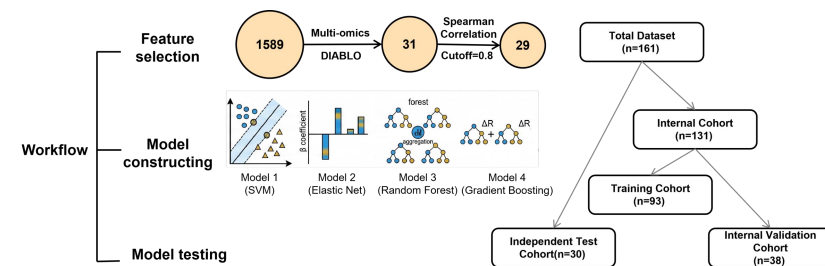


Figure 1. Workflow of the predictive model construction. The pipeline illustrates the three main stages: feature selection from baseline multi-omics data, construction of four machine learning models, and dataset allocation (total n=161) into training, internal validation, and independent test cohorts for comprehensive model testing.

Result

Four machine learning models, including Elastic Net (EN), Radial-kernel Support Vector Machine (SVM), Gradient Boosting Machine (GBM) and Random Forest (RF), were trained and validated. The SVM model outperformed other models, achieving AUCs of 0.763 and 0.884 in the validation and independent test cohorts, respectively. The Elastic Net model demonstrated robust stability with AUCs of 0.798 and 0.792 (Figure 2).

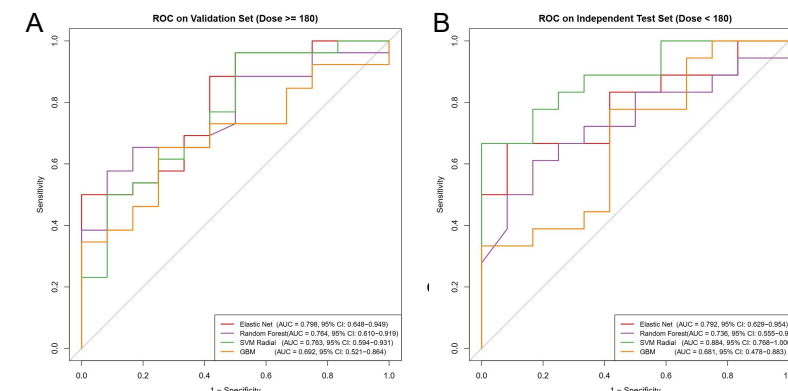


Figure 2. Predictive performance of machine learning models for therapeutic response.

Risk scores were calculated for each patient, and patients were stratified into high-risk and low-risk groups based on the median risk score. The median PFS of low-risk patients was significantly extended to approximately 22.0 months, compared to 9.5 months for high-risk patients (log-rank $p < 0.0001$). Similarly, the median OS for the low-risk cohort was notably prolonged (~33.0 months) versus the high-risk cohort (~18.0 months) (log-rank $p < 0.0001$).

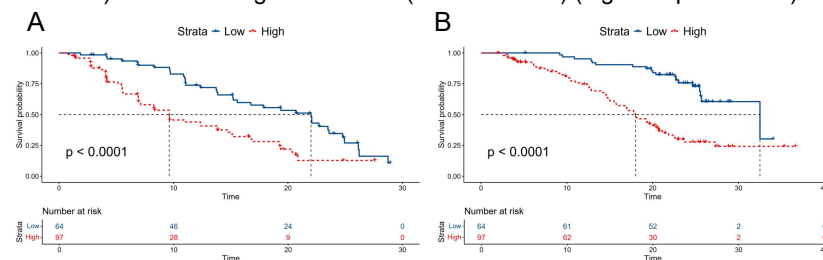


Figure 3. Kaplan-Meier survival analysis for progression-free survival (PFS) and overall survival (OS). Patients were stratified into high-risk and low-risk groups based on the median risk score derived from the baseline multi-omics model.

Result

A 29-marker signature (20 proteins, 9 metabolites) was identified, which was significantly enriched in blood coagulation and integrin signalling pathways. Moreover, key hub molecules identified via Elastic Net weighting showed a strong correlation with both therapeutic response. (Figure 4).

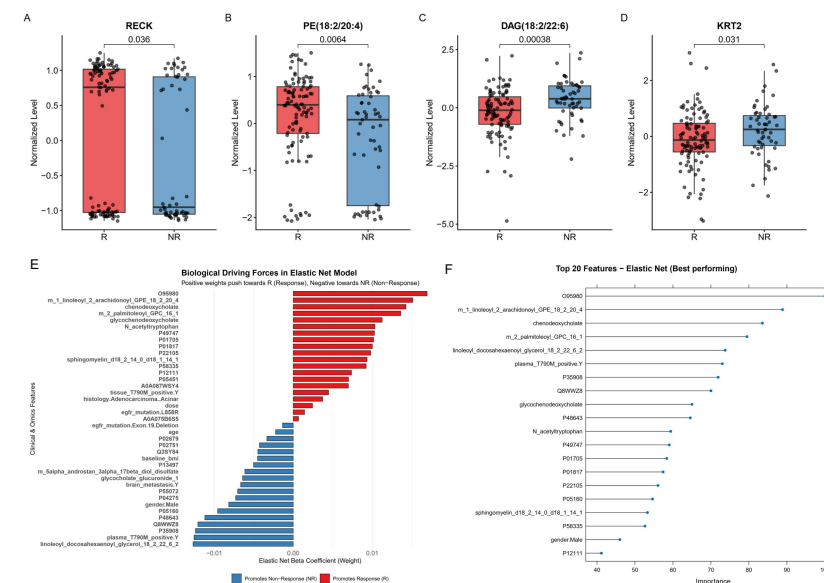


Figure 4. Identification of key hub molecules and biological pathway enrichment from the baseline multi-omics signature.

Conclusion

The integration of proteomics and metabolomics via machine learning framework identified 29-molecule signatures for risk stratification system constration, which can potentially predict third-generation EGFR-TKI efficacy in a non-invasive way for NSCLC patients, complementing current clinical evaluations.