正文

In order to identify diagnostic biomarkers capable of distinguishing between patients with high and low overall survival (OS) following chemotherapy, feature selection was conducted on a dataset comprising 171 metabolites extracted from tissue samples. Utilizing the selected metabolites, four predictive models based on the training dataset were fitted and then assessed the independent performance of both the models and the feature selection technique using a separate testing dataset (Supplementary Figure). Accounting for the complexity inherent in clinical data and the characteristics of metabolites, the XGBoost algorithm exhibited superior performance compared to the other models, making it the preferred choice due to its effectiveness in handling complex feature interactions and its potential for high predictive accuracy. The eXtreme Gradient Boosting Classifier (XGBoost), employing a panel of 6 metabolite biomarkers, effectively differentiated between low-OS and high-OS patient groups, achieving an area under the receiver operating characteristic curve of 0.8996 (0.6667–1.0000). These results collectively highlight significant metabolites alterations between the tissues of patients with varying survival risks and underscore the promise of metabolites as biomarkers for enhancing the diagnosis of small cell lung cancer (SCLC).

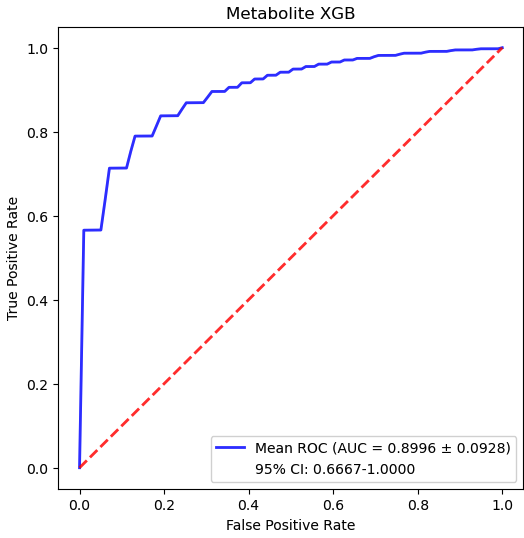
Methods--Machine learning for the diagnosis of SCLC and prediction of survival outcomes

```methods部分和lipid是一样的```

To distinguish survival risks, we first identified statistical differences using the Mann-Whitney U-test to select a reduced number of diagnostic features for classification on the training dataset. Then five-fold cross-validation was conducted on the training set to obtain stable feature selection results. A P-value threshold of 0.05 was set and we selected features showing significant differences. For the prediction of survival outcomes, 76 participants who underwent chemotherapy were divided into high overall survival rate and low overall survival rate groups based on a commonly used threshold of 24 months (2 years). The patients were then randomly divided into training and testing sets. Features were more strictly selected based on their variable importance, as represented by Shapley values, with a mean Shapley value rank of 20 required as the threshold.

Once feature selection was completed, a prediction model was built using the eXtreme Gradient Boosting (XGBoost) method based on the training dataset. This diagnostic and prediction model was then applied to diagnose patients and predict survival outcomes in the testing set. The XGBoost model was configured with a max\_depth of 6 to balance model complexity and overfitting, a learning rate of 0.1 for gradual updates, and 1000 estimators for sufficient model capacity. XGBoost was performed using the xgboost package (v.2.0.3) and SHAP analysis was performed using the shap package (v.0.42.1) in Python (v.3.9.18).

Figure

1. Receiver operating characteristic (ROC) curve of the XGBoost model in the testing set.  
   
2. Schematic of the dataset creation and analysis strategy for the distinguish of Survival Rate

