Machine learning to predict lymph node metastasis in T1 esophageal squamous cell carcinoma：A multicenter study

July 15，2021

Statistical Analysis Plan

Predictor characteristics

Machine-learning models were developed using data on patients' age and sex, depth of tumor invasion, tumor size, tumor location, macroscopic tumor type, lymph vascular invasion, and histologic grade. Most of our data were structured as binary features except for age and tumor size. Location was three categorical variables. Age and tumor size were normalized to be in a range [0, 1]. The Fisher exact test and the Kolmogorov-Smirnov test were used for assessing categorical variables or continuous variables between groups. Statistical significance was set at $p<0.05$.

ML model building and validation

Elastic net regularized logistic regression (ELR) and random forest (RF) and extreme gradient boosting (XGB) were applied to predict LNM.ELR mixes the penalties of ridge and lasso to minimize the overfitting and help variable selection. RF combines a predetermined number of decision trees (usually about 1000) generated on a random subgroup of the data set. XGB aims to improve consecutively. XGB aims to improve consecutively by creating models to explain where the previous model fails and then repeating this process (usually around 1000 times). At the same time, regularization is used to reduce overfitting. The individual models were combined to generate overall predictions. This strategy, theoretically, is advantageous when utilizing a variety of model types that represent various aspects of patients' risk profiles.

Random Over Sampling Examples（ROSE）was applied to deal with the class imbalance problem. During model creation, a 10-fold cross-validation with five repeats used for each model's hyperparameter tuning. And log loss was used as optimization metric. For ELR, the hyperparameter tuning was conducted for α and λ hyperparameters. 1000 decision trees were used to create the RF model. The split rule, minimum node size, and number of variables per tree were all hyperparameter

tuned. THE hyperparameters for the XGB model include maximum tree depth, number of optimization rounds, minimum weight in each child node, minimum loss reduction (γ), regularization penalty (η), and subsampling for regularization. These three best-performed models were then integrated to create the ensemble model（Ens） by applying logistic regression to create a linear blend of projected probabilities.

The area under the receiver operator characteristic (ROC) curve (AUC) was used to evaluate the models' discrimination power. Internal validation was carried out with 1000 resampled data sets and 0.632 bootstrapping strategy. Calibration was assessed visually and formally with the Hosmer–Lemeshow test. Isotonic regression was used to scale probabilities on the final model to enable meaningful interpretation of probability.

The VarImp function of the caret R package was used. The contribution of each variable to the global ROC curve was calculated as a percentage.

Sub-analysis

The diagnostic ability of the ML model was compared with the NCCN guidelines for LNM. In the NCCN guideline, poorly differentiated tumors, deep submucosal invasion, and lymphovascular invasion are considered predictive of LNM.

Data analysis was conducted using R version 3.5.3 (R Foundation for Statistical Computing, Vienna, Austria). The caret and caretEnsemble packages were used to train the models. The full R code to train the models is available supplementary, along with a list of packages used.

The calibrated final model was designed using R Shiny (available freely at https://predicted.shinyapps.io/Rshinyfinal/). No data entered into the model were collected or stored.