Establishment of prediction models for lung cancer NOG/PDX models: A guideline for machine learning in small biomedical datasets

Haoyue Guo
Department of Medical Oncology, Shanghai Pulmonary Hospital, Tongji University Medical School Cancer Institute, Tongji University School of Medicine, No 507 Zhengmin Road, Shanghai 200433, China. School of Medicine, Tongji University, No 1239 Siping Road,

Li Diao
Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, No 800 Dongchuan Road, Shanghai 200240, China

Hui Qi
Oncology and immunology BU, Research service division, WuXi Apptec, Shanghai 200131, China

Chunlei Dai
Oncology and immunology BU, Research service division, WuXi Apptec, Shanghai 200131, China

Yu Chen
Spine center, Orthopedic department, Shanghai Changzheng Hospital, No 415, Fengyang Road, Shanghai 200003, China

Xuzhen Tang
Oncology and immunology BU, Research service division, WuXi Apptec, Shanghai 200131, China

Yayi He (2250601@qq.com)
Tongji University Affiliated Shanghai Pulmonary Hospital

Research

Keywords: Lung cancer, Patient-derived tumor xenografts, NOG mice, Machine learning

Posted Date: September 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-73446/v1

Abstract

Background: Targeted therapy and immune checkpoint inhibitors are the most promising treatments for lung cancers but still facing multiple challenges, including resistance and individual difference. Therefore, patient-derived tumor xenografts (PDX) models are developed for drug discovery and screening. NOG mice is under the destruction of the interleukin-2 (IL-2) receptor common gamma chain, which is appropriate for building PDX models to test immunotherapies. However, current studies have little understanding of the causes of genotype mismatches in PDX or NOG/PDX models, which leads to a massive economic and time loss.

Methods: Lung cancer tissues from 53 patients were obtained and engrafted into NOG mice. All of the patients' tumors and NOG/PDX models were detected for common gene mutations. Seventeen clinicopathological features were organized and input to stepwise logistic regression based on the lowest Akaike information criterion (AIC), least absolute shrinkage and selection operator (LASSO)-logistic regression, support vector machine recursive feature elimination (SVM-RFE), eXtreme Gradient Boosting (XGBoost), Gradient Boosting & Categorical Features (CatBoost), and synthetic minority over-sampling technique (SMOTE). Finally, the performance of all models was evaluated by the accuracy, area under the receiver operating characteristic curve (AUC), and F1 score in 100 testing groups.

Results: Fifty-three lung cancer NOG/PDX models were successfully established, with a genotype matching rate of 79.2% (42/53). Two multivariable logistic regressions revealed that age, the number of driver mutations, epidermal growth factor receptor (EGFR) gene mutations, the type of prior chemotherapy, prior tyrosine kinase inhibitors (TKIs) therapy, and the source were potent predictors. Moreover, CatBoost (*mean accuracy=0.960; mean AUC=0.939; mean F1 score=0.908*) and 8-feature SVM (*mean accuracy=0.950; mean AUC=0.934; mean F1 score=0.903*) showed the best performance compared with the other algorithms. Moreover, the combination of SMOTE with SVM significantly improved the predictive capability (mean accuracy: *0.961 vs. 0.958, P=0.025*; mean AUC: *0.940 vs. 0.935, P=0.045*; mean F1 score: *0.909 vs. 0.903, P=0.047*).

Conclusions: We established an optimal predictive model to screen lung cancer patients for NOG/PDX models, and also offered a general approach for building prediction models in small unbalanced biomedical samples.

Background

Lung cancer is the one that causes most deaths in humans, the number of which exceeds one million each year worldwide[1]. Around 85% of lung cancers are non-small cell lung cancer (NSCLC), and small cell lung cancer (SCLC) accounts for 15% of lung cancers[2]. Recently, the chemotherapy-based paradigm in lung cancer patients has been shifted with the introduction of the driver genes and the advance of molecular detection technology[3, 4], especially in those with epidermal growth factor receptor (EGFR) positive mutant[5, 6] and rearrangements of the anaplastic large-cell lymphoma kinase (ALK) genes[7].

Nevertheless, targeted therapy is facing a series of difficulties, including different individual responses and frequently acquired resistance[8, 9]. Subsequently, immune checkpoint inhibitors (ICIs) are recommended as substitutional drugs of targeted therapy[10]. Therefore, the screening of immunotherapy for patients without driver gene mutations or response to targeted drugs is pivotal to refractory lung cancers.

To overcome these challenges, pre-clinical animal models are crucial for drug screening and efficacy evaluation to achieve precision medicine. Patient-derived tumor xenografts (PDX) have emerged as an accurate pre-clinical system capable of maintaining the molecular, genetic, and histopathologic heterogeneity of the parental tumors[11, 12]. Moreover, a new generation of super immunodeficient mice named NOG was considered as an excellent choice to build PDX models for cancer immunotherapy, with the destruction of the interleukin-2 (IL-2) receptor common gamma chain and functional deficiency of multiple immune cells such as T cells, B cells, natural killer cells (NK cells), macrophages, and dendritic cells[13]. Compared with nude mice and traditional severe combined immunodeficient (SCID) mice, NOG mice have demonstrated great potential for researches on ICIs, adoptive T-cell therapy (ACT), and other immunotherapy since tumor-infiltrating T lymphocytes (TILs) can be serially transplanted into them after xenografts develop[14, 15].

Nonetheless, the low success rate of PDX establishment (20–40%)[16–18] with the notably inconsistent rate of diver gene-mutations between established PDX and original tumors (10–20%)[19, 20] is quite a worrisome problem, but not well studied. Since protocols of PDX models are time-consuming, labor-intensive, and costly[21], the presence of inconsistency of driver gene mutations imparts a vast toll on researchers, physicians, and patients. Although multiple factors including gender, smoking history, pathology, TNM-stage, tumor grade, the quality of tumor samples, and EGFR gene mutations have been assumed to correlate with the success rate of the tumor engraftment[17, 18, 22], whether these factors contribute to the consistency of driver gene mutations PDX models, especially those established in NOG mice have not been validated. This study is aimed to perform machine learning (ML) algorithms including multivariable logistic regression (LR), support vector machine recursive feature elimination (SVM-RFE), gradient boosting decision tree (GBDT), and synthetic minority over-sampling technique (SMOTE) to establish a potent tool for predicting inconsistency of driver gene mutations between NOG/PDX models and patients' tumors (Fig. 1).

Methods

Patient samples

Lung cancer tissues or cells from 53 patients were obtained from computed tomography (CT) guided percutaneous lung biopsy (CT-PLB), lymph node biopsy (LNB) or thoracentesis at Shanghai Pulmonary Hospital (Shanghai, China), between August 2018 and October 2019.

Preparations for Tissues samples

Harvested tissues from TBB, CT-PLB, and LNB were divided into three parts. The first part was minced into fragments of 50–100 mm^3 and was immersed in frozen medium Bambanker (Nippon Genetics, Cat. No. BBH01), and then was kept in liquid nitrogen until implantation into immunodeficient mice. The second part was immediately frozen in liquid nitrogen for DNA/RNA extraction. The third part was made into formalin-fixed paraffin-embedded (FFPE) slides for pathologic assessment.

Preparations for the malignant pleural effusion (MPE)

The preparations and culture of MPE were conducted as previously described. Approximately 200–1000 ml of pleural effusion was extracted each time by thoracentesis. The samples were centrifuged at 450 G for 10 min and then were resuspended in PBS. The tumor cells were isolated from the interphase layer of samples by density gradient centrifugation using Ficoll-PaqueTM PLUS (GE Healthcare Bio-Sciences, Uppsala, Sweden). After being washed with PBS, the tumor cells were cultured in RPMI-1640 containing 10% fetal bovine serum (FBS) and 10 ng/ml epidermal growth factor (EGF) at a density of 1 to 2×10^6 cells per plate.

NOG/PDX establishment

All animal experiments in this study are following the guidelines of the Institutional Animal Care and Use Committee (IACUC). The PDX models were established in 6-8-week-old female NOG mice (NOD.Cg-$Prkdc^{scid}IL2rg^{tm1Sug}$/JicCrl) (Charles River, Beijing, China).

Frozen tissues were thawed at 37 °C and directly implanted into the sterilized skin on NOG mice subcutaneously (n = 4–5 for each tumor sample). Simultaneously, tumor cells isolated from MPE mentioned above were washed once in PBS and then injected 510^6 cells into the right flank of each NOG mice (n = 4–5 for each MPE sample).

The initial tumor-implanted NOG mice were maintained for 120 days and were measured once a week. The tumor volumes (TV) were measured by the formula: TV= (length × $width^2$)/2 (length was the longest diameter, while the width was the shortest diameter). The xenografted tumor was passaged when the tumor size reached around 700–800 mm^3, and the PDX models utilized in this study were from the third passage (P3) to the fifth passage (P5). The PDX tumors of each passage were again separated into three parts. The first part was implanted into another NOG mouse for passaging. The second part was immediately frozen in liquid nitrogen for DNA/RNA extraction. The third part was made into FFPE slides for pathologic assessment.

DNA and RNA extractions

Lung cancer tissues and PDX tissues were pathologically reviewed to ensure that tumor cells were more than 80% and that no significant tumor necrosis had occurred before the DNA extraction. Genomic DNA was extracted from each tissue sample using QIAamp DNA Mini Kit (Qiagen-51306, Germany). The quantity and purity of DNA samples were measured using Nanodrop ND-1000 UV/VIS Spectrophotometer (Therm Scientific, USA). DNA fragment integrity was confirmed by electrophoresis using 1% agarose gel.

The concentration of DNA samples was normalized to 20 ng/μL and stored at − 20 °C until use. Both ‘Hot spot’ mutations in EGFR (exon 18, 19, 20, 21) and ALK fusions (EML4-ALK) were screened by amplification refractory mutation system (ARMS) and mutant-enriched liquid chip polymerase chain reaction (PCR) method.

Models' performance evaluation

To evaluate the performance of all prediction models in this study, we calculated the following indexes:

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

(1) Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$, (2)Precision=$\frac{TP}{TP+FP}$, (3)Recall=$\frac{TP}{TP+FN}$, (4)F1 score=$\frac{2\times Precision\times Recall}{Precision+Recall}$, (5) Receiver operating characteristic curve (ROC) takes false positive rate as the abscissa and true positive as the ordinate. The area under the ROC (AUC) was calculated by “sklearn.metrics” of Python.

Statistical analysis

Stepwise LR based on the lowest AIC, logistic-LASSO regression, SVM-RFE, eXtreme Gradient Boosting (XGBoost), and Gradient Boosting & Categorical Features (CatBoost) were utilized for developing multivariate prediction models (detailed information was presented in the Supplementary Methods). Stratified random sampling was performed to generate 100 training groups (containing seven non-matching samples and 28 matching samples) and 100 testing groups (containing four non-matching samples and 14 matching samples). The pair-sample t-test was used to compare the performance among different models in testing groups. All of the data analysis in this study was performed by SPSS (version 23.0, IBM SPSS), R software (version 3.1.0, R Core Team), Matlab (version 7.12.0, Mathworks), and Python (version 2.7, Python Software Foundation). All of the Figures were produced via GraphPad Prism (version 8.0, GraphPad Software). Statistical tests were two-sided, and $P < 0.05$ was considered statistically significant.

Results

Establishment of NOG/PDX models

The general clinicopathologic of all 53 NSCLC patients were shown in Table S1. The median patient age was 66 years, and 83.2% (44/53) were male. Three (6%) patients were diagnosed as TNM-stage I and the others were diagnosed as stage III/IV (94%). Forty patients (75.5%) were diagnosed as NSCLC, including 15 squamous cell carcinomas (SCC), nine adenocarcinomas (ADC), and six other NSCLCs, while thirteen (24.5%) patients were SCLC. Among all of the samples, tissues from 10 patients (18.9%) were with EGFR

gene mutations, one patient (1.9%) was with ALK infusions, and the rest 42 samples were non-mutant tissues (79.2%). There were 39 patients (73.6%) with metastatic sites. Forty-eight samples (90.6%) were obtained by CT-PLB, while LNB obtained two samples (3.8%), and three samples (5.7%) were obtained by thoracentesis. There were 39 cases (73.6%) receiving therapy before sampling, including chemotherapy (n = 29), tyrosine kinase inhibitors (TKIs) (n = 8) and immunotherapy (n = 2), and 14 cases (26.4%) without any prior therapy.

All of the PDX models included in this study were all confirmed as the successful establishment (size reached around 700–800 mm^3) by pathologists, and the representative sections of PDX models with haemotoxylin and eosin (H&E) staining were shown in Fig. 1. The overall rate of the driver genes matching was 84% (42/50).

Model 1: Logistic Regression

1 Univariable analysis of factors associated with genotype mismatches

Logistic univariate regression analysis indicated that the risk factors for the inconsistency of driver genes mutations between PDX models and parental tumors were female, younger age, smoking history, acquisition from LNB or thoracentesis, NSCLC except for squamous cell carcinoma (SCC), EGFR mutations, more driver genes mutations, without prior chemotherapy, with prior chemotherapy of pemetrexed plus carboplatin, and prior TKIs therapy (Table 1).

Table 1
The characteristics of 53 patients and univariable logistic regression of 17 clinicopathological variables for determining factors correlated to the inconsistency of driver genes between PDX models and parental tumors

	Matching of driver genes between PDX models and parental tumors		P†	OR (95% CI) †
Variables	**Yes**	**No**		
Mean Age (years)	65.36	59.73	**0.05**	0.921 (0.848-1.000)
Gender (n)			**0.01**	
Male	39	5		
Female	3	6		15.600 (2.938–82.836)
Smoking status (n)			**0.01**	
No	12	9		
Yes	30	2		11.250 (2.113–59.884)
Source (n)			**0.042**	
CT-PLB	40	8		
LNB or thoracentesis	2	3		7.5 (1.704–52.377)
Pathology (n)			**0.012**	
ADC	8	7		10.500 (1.076-102.478)
SCC	9	0		0
Other NSCLCs	13	3		2.769 (0.252–30.383)
SCLC	12	1		1
EGFR mutation (n)			**< 0.001**	
No	41	2		

† The P value and odds ratio are analyzed by univariate logistic regression. CT-PLB: computed tomography-guided percutaneous lung biopsy; LNB: lymph node biopsy; SCC: squamous cell carcinomas; ADC: adenocarcinomas; NSCLC: non-small cell lung cancer; SCLC: small cell lung cancer; EGFR: epidermal growth factor receptor; EC: etoposide and carboplatin; GC: gemcitabine and cisplatin; AC: pemetrexed and carboplatin; PR: partial response; PD: progressive disease; SD: stable disease.

	Matching of driver genes between PDX models and parental tumors		P†	OR (95% CI) †
Variables	**Yes**	**No**		
Yes	1	9		184.500 (15.046-2262.404)
The number mutations (n)			**< 0.001**	
0	41	1		
1	1	7		287.000 (16.024-5140.254)
2	0	3		6.62E + 10
T-stage (n)			0.148	
1–2	12	1		
3–4	30	10		4.000 (0.460–34.750)
N-stage (n)			1	
0	5	1		
1	4	0		0.700 (0.068–7.201)
2	12	4		0
3	21	0		1.167 (0.274–4.976)
M-stage (n)			0.173	
0	13	1		
1	29	10		0.223 (0.026–1.929)
TNM-Stage (n)			1	
1	3	0		
3	11	1		0

† The P value and odds ratio are analyzed by univariate logistic regression. CT-PLB: computed tomography-guided percutaneous lung biopsy; LNB: lymph node biopsy; SCC: squamous cell carcinomas; ADC: adenocarcinomas; NSCLC: non-small cell lung cancer; SCLC: small cell lung cancer; EGFR: epidermal growth factor receptor; EC: etoposide and carboplatin; GC: gemcitabine and cisplatin; AC: pemetrexed and carboplatin; PR: partial response; PD: progressive disease; SD: stable disease.

	Matching of driver genes between PDX models and parental tumors		P†	OR (95% CI) †
Variables	**Yes**	**No**		
4	28	10		0.255 (0.029–2.231)
Number distant metastatic sites (n)			1	
0	13	1		
1	17	5		0.077 (0.002–2.394)
2	5	2		0.294 (0.015–5.595)
3	5	1		0.400 (0.016–10.017)
4	1	1		0.200 (0.006–6.664)
5	1	1		1 (0.020-50.397)
Prior therapy (n)			0.473	
No	12	2		
Yes	30	9		1.800 (0.338–9.581)
Prior chemotherapy (n)			**0.005**	
No	15	9		8.100 (1.545–42.476)
Yes	27	2		
Chemotherapy type (n)			**0.037**	
EC	7	0		
GC	7	0		1
Paclitaxel liposome	3	0		1
AC	3	1		538491658.7

† The P value and odds ratio are analyzed by univariate logistic regression. CT-PLB: computed tomography-guided percutaneous lung biopsy; LNB: lymph node biopsy; SCC: squamous cell carcinomas; ADC: adenocarcinomas; NSCLC: non-small cell lung cancer; SCLC: small cell lung cancer; EGFR: epidermal growth factor receptor; EC: etoposide and carboplatin; GC: gemcitabine and cisplatin; AC: pemetrexed and carboplatin; PR: partial response; PD: progressive disease; SD: stable disease.

	Matching of driver genes between PDX models and parental tumors		P†	OR (95% CI) †
Variables	**Yes**	**No**		
Other chemotherapy	7	1		2307282139.4
None	15	9		989284985.7
Prior TKIs therapy (n)			**0.005**	
No	39	6		
Yes	5	3		10.833 (2.040-57.525)
Curative effect of prior therapy (n)			0.387	
No therapy	12	2		
PR	3	0		0.417 (0.076–2.296)
PD	3	0		0
SD	4	1		0
Not evaluated	20	8		0.625 (0.060–6.486)

† The P value and odds ratio are analyzed by univariate logistic regression. CT-PLB: computed tomography-guided percutaneous lung biopsy; LNB: lymph node biopsy; SCC: squamous cell carcinomas; ADC: adenocarcinomas; NSCLC: non-small cell lung cancer; SCLC: small cell lung cancer; EGFR: epidermal growth factor receptor; EC: etoposide and carboplatin; GC: gemcitabine and cisplatin; AC: pemetrexed and carboplatin; PR: partial response; PD: progressive disease; SD: stable disease.

2 Multivariable selection in all 53 examples

2.1 Logistic regression based on the Akaike information criterion (AIC)

To balance the prediction model's performance and complexity, we performed stepwise model selections by calculating the AIC. According to the univariable analysis, there are ten potential predictive features. Figure 2A showed the AIC values of each step in the backward stepwise LR, where ten predictive features were deleted one by one until AIC could not decrease any more. Generally, models excluding the number of driver genes presented the worst AIC, which indicated that the number of driver genes was a weighty predictor. Moreover, the best multivariable models selected by AIC was a five-variable logistic regression including age, the number of driver mutations, type of prior chemotherapy, prior TKIs therapy, and the source.

2.2 Least absolute shrinkage and selection operator (LASSO)-logistic regression

We performed the LASSO regularization on the LR to improve the prediction accuracy and interpretability. Two features out of the ten features were screened out via a LASSO logistic regression combined with ten-fold cross-validation, where the optimal penalization coefficient λ valued as one standard error: EGFR mutations and the number of driver genes (Figs. 2B and 2C).

Model 2: Support vector machine recursive feature elimination (SVM-RFE)

SVM-RFE begins with a complete feature set and eliminates the least important feature for classification in each iteration, according to the weight vector of dimension length. According to the rank of feature importance, which was visualized in Fig. 3, we firstly deleted the least important seven variables and then eliminated the remaining ten variables one by one to optimize the prediction accuracy. According to the mean predictive accuracy and F1 score in 100 testing groups by stratified random sampling, the SVM model, including eight variables, maintained the best performance with the least complexity. As a result, the 8-feature SVM was the most optimal model among all SVM classifiers.

Model 3: Gradient Boosting Decision Tree (GBDT)

To implement GBDT, we performed two commonly used algorithms, eXtreme Gradient Boosting (XGBoost) and Gradient Boosting & Categorical Features (CatBoost). A large number of experiments indicated that multicollinearity among features did not hinder decision trees' predictive classifications [23]. Therefore, we input all of the 17 features to XGBoost and CatBoost in this study. The rank of features based on XGBoost and CatBoost classification algorithms was also presented in Fig. 3. The representative structure of a decision tree generated by XGBoost and CatBoost was shown in Fig. 4B.

Modeling in training groups and evaluating performance in testing groups

(1) Comparison between different models

According to AUC, accuracy and F1 score of 100 testing groups, CatBoost (*mean accuracy = 0.960; mean AUC = 0.939; mean F1 score = 0.908*) and 8-feature SVM (*mean accuracy = 0.950; mean AUC = 0.934; mean F1 score = 0.903*) dramatically exceeded the other three models, XGBoost(*mean accuracy = 0.951; mean AUC = 0.908; mean F1 score = 0.873)*, LASSO-LR (*mean accuracy = 0.937; mean AUC = 0.886; mean F1 score = 0.841*), and LR based on AIC (*mean accuracy = 0.923; mean AUC = 0.850; mean F1 score = 0.789*). In spite that the accuracy of 8-feature SVM and XGBoost were statistically equal, CatBoost and 8-feature SVM indicated the overall best performance. Additionally, the mean accuracy (*P = 0.103*), AUC (*P*

= 0.066), and F1 score (*P = 0.128*) between CatBoost and 8-feature SVM were not significantly different (Fig. 5A), promisingly breaking the limitation of the imbalanced small sample dataset.

(2) The improvement upon the synthetic minority over-sampling technique (SMOTE) application

Synthetic minority oversampling technique (SMOTE) is a kind of oversampling method that increases the number of positive classes through randomly data replication to achieve an equal number with the negative class[24]. Herein we exerted SMOTE to add ten more positive samples into training groups to establish every model and then tested them in the original 100 testing groups (Table S2). After performing the paired-sample t-test, 8-feature SVM combined with SMOTE achieved a significant improvement on accuracy (*0.961 vs. 0.958, P = 0.025*), AUC (*0.940 vs. 0.935, P = 0.045*), and F1 score (*0.909 vs. 0.903, P = 0.047*) (Fig. 5B). Among all the other algorithms, XGBoost was the only model that acquired an obvious advantage in AUC (*0.935 vs. 0.908, P = 0.004*; Fig. 5C), while the other models did not present any improvement (Table S3). Therefore, we exhibited an approach that can improve the performance of SVM in small uneven samples.

Discussion

Overall, this study initially developed a prediction model for the inconsistency of driver gene mutations between NOG/PDX models and patients' samples. A total of 53 lung cancer NOG/PDX models were successfully engrafted and excised, including 42 NOG-PDX models with matching driver gene mutations from parents' tumors and 11 NOG-PDX models with non-matching ones. To analyze this small unbalanced database, we performed three models of five algorithms, including LR based on AIC, LASSO-LR, SVM, XGBoost, and CatBoost, which all present with an excellent predictive capability. According to the evaluation indexes in testing groups, CatBoost and SVM demonstrated the best performance of and modeling. Moreover, the application of SMOTE generally improved the performance of SVM based on the fundamental level.

LR illustrated what exactly determined the genotypes of NOG/PDX models

LR is a widely ML technique in biomedical data analysis since it is reasonably easy to interpret with a clear demonstration of the positive or negative association of the variables with the predicted probability[25]. Therefore, the two multiple LR models revealed the critical predictor variables.

The formation and passaging of PDX models are dynamic events, where clonal and subclonal alternations frequently occur, especially when the development of P1 PDX models is slow, which gives adequate time for tumor cells to mutate for adaptation a new environment[26, 27]. In addition to these cell-autonomous heterogeneities, the stromal heterogeneity in the tumor microenvironment (TME) is another critical reason for different driver genotype of PDX from parental tumors[12]. As expected, most

of the predictive features from univariable analysis consist of the factors associated with xenografts engraftment, except for the pathology. It was reported that SCC was much more prone to be tumorigenic in nude mice compared to adenocarcinoma ADC [18], which is contrary to our conclusion that SCC is the most challenging type to establish NOG/PDX models of genetic matching. More $CD8^+$ TILs were detected in the cancer nests of SCC than in non-SCC [28], which reveals that PDX models of SCC may lose more tumor stroma during the xenograft engraftment. Moreover, SCC was prone to carry significantly more clonal mutations than AC[29], which may also contribute to more clonal selections.

As for multivariable analysis, age, the number of driver mutations, EGFR mutations, the type of prior chemotherapy, prior TKIs therapy, and the source shown a significant role in the inconsistency of driver gene mutations. Although age accounted for a small weight in multivariable LR, we have not found an appropriate therapy to illustrate a younger age rather than an elder age is a risk factor for the inconsistency of driver gene matching[30]. However, this surprising factor suggests that the age of implemented mice might play a critical role in the establishment of PDX models. Most of the PDX models in current researches use 8-week-old mice rather than aged mice (> 8 months). Recent studies found that aging could dramatically alter the components of the tumor microenvironment[31], thereby the inconsistent age of mice from patients could be the potential reason why age becomes a predictive feature here. Another feature, source, also played a negative role in matching the genotype, different from that in tumor engraftments. Although fluids source, including MPE and lymph, are proved to have a higher engraftment rate than the solid tumor tissues[32], we found that fluid-derived tumor xenografts were more challenging to maintain driver genotypes from parental tumors.

The number of driver gene mutations, including clonal and subclonal mutations, are associated with intra-tumor heterogeneity, genomic instability, or chromosomal instability[33]. The largest coefficient of the number of the driver gene in the multivariable LR model also illustrates its absolute importance in developing non-patient-matched genotypes. Secondly, PDX models from EGFR mutant lung cancers were reported with poor histological differentiation, and frequent loss of EGFR mutations[34], which supported the high inconsistent risk of EGFR mutant NOG/PDX models in this study. Thirdly, the evidence that pemetrexed increased the number of TILs, and upregulated immune-related genes related to antigen presentation might support the conclusion that PDX models from patients receiving pemetrexed are less likely to maintain the original genotypes[35]. TKIs have been proved with the capability to alter the pulmonary TME, including increased $CD8^+$ T cells and mononuclear myeloid-derived suppressor cells (M-MDSCs) ($CD11b^+Ly6^-G^-Ly6C^{high}$), and fewer $Foxp3^+$ T regulatory cells (Tregs) and M2-like macrophages ($CD206^+$)[36]. Also, the clonal selection is a frequent occurrence during TKIs therapy, resulting in TKI resistance[37]. Interestingly, we found that the factors promoting TILs were conducive to the stability of genotypes during the NOG/PDX models establishment, which needs further verification (Fig. 6).

SVM-RFE and GBDT provide a robust and straightforward classifier

Unlike LR, both SVM and GBDT are similar to "black boxes," which only shows the inputs and outputs without internal workings[38]. SVM and GBDT are considered with the reliable power for classification, less concern for overfitting, and the ability to handle unbalanced data, which has been validated in this study. Thereby, when there is no need to explain the model in detail with an immediate requirement of building an accurate classifier, CatBoost, SVM, or SVM-SMOTE become a better choice for predicting the inconsistency of driver gene mutations with a significantly better performance.

ML for small biomedical unbalanced datasets

Recently, ML is a promising topic for predictive modeling in numerous areas, which enables prediction models to “learn” information systematically from initial data and adapt to each new data environment[39]. However, ML has not been widely performed in small sample databases (less than ten frequencies per predictor variable), which is a common characteristic in biomedical animal models with expensive costs and complicated techniques[40]. Ultimately, the ML algorithms we attempted to establish predictive tools for lung cancer NOG/PDX models demonstrated excellent performance, which not only provides a predictive tool to screen lung cancer patients for NOG/PDX models of precisive immunotherapy but also offers a general approach for building prediction models in small biomedical samples:

(1) Select features to develop a multivariable model in all samples with standard ML algorithms, including stepwise LR based on AIC, LASSO-LR, SVM (or SVM-SMOTE), XGBoost, CatBoost, and et al.

(2) Perform stratified random sampling to generate 100 training groups and testing groups to achieve stable performance.

(3) Formulate the predictive score or establish the predictive classifier in training groups.

(4) Evaluate the predictive model based on ROC, accuracy, and F1 score, in the corresponding testing groups to determine an optimal algorithm or modeling.

(5) Interpret the critical predictors for positive class by LR, and apply the optimal algorithm for the final prediction.

Conclusions

In conclusion, we established a predictive model for the inconsistency of driver gene mutations between NOG/PDX models and patients' samples based on machine learning, which promises to improve the success rate of PDX establishment and reduce the massive economic loss. Further, the NOG mice we used in this study were considered an excellent choice to build PDX models for cancer immunotherapy, but not well studied. Therefore, the model we established has the potential for immunotherapy screening and development. What is more, machine learning has not been widely performed in small sample databases. Ultimately, the algorithms we explored in this study also offer a general approach for building prediction models in small biomedical samples.

Abbreviations

ACT
adoptive T-cell therapy
ADC
adenocarcinomas
AIC
akaike information criterion
ALK
anaplastic large-cell lymphoma kinase
ARMS
amplification refractory mutation system
AUC
the area under the receiver operating characteristic curve
CatBoost
gradient boosting & categorical features
CT
computed tomography
CT-PLB
computed tomography guided percutaneous lung biopsy
EGF
epidermal growth factor
EGFR
epidermal growth factor receptor
FBS
fetal bovine serum
FFPE
formalin-fixed paraffin-embedded
GBDT
gradient boosting decision tree
H&E
haemotoxylin and eosin
IACUC
Institutional Animal Care and Use Committee
ICIs
immune checkpoint inhibitors
IL-2
interleukin-2
LASSO
least absolute shrinkage and selection operator

LNB
lymph node biopsy
LR
logistic regression
ML
machine learning
M-MDSCs
mononuclear myeloid-derived suppressor cells
MPE
malignant pleural effusion
NK cells
natural killer cells
NSCLC
non-small cell lung cancer
PCR
polymerase chain reaction
PDX
patient-derived tumor xenografts
ROC
receiver operating characteristic curve
SCC
squamous cell carcinomas
SCID
severe combined immunodeficient
SCLC
small cell lung cancer
SMOTE
synthetic minority over-sampling technique
SVM-RFE
support vector machine recursive feature elimination
TILs
tumor-infiltrating T lymphocytes
TKIs
tyrosine kinase inhibitors
TME
tumor microenvironment
TV
tumor volumes
XGBoost
eXtreme Gradient Boosting

Declarations

Ethics approval and consent to participate: The study was approved by the ethics/licensing committee of Shanghai Pulmonary Hospital (approval number: NO K18-203).

Consent for publication: The authors have obtained consent to publish from the participants to report individual patient data.

Availability of data and materials: All data generated or analyzed during this study are included in this published article or supplementary material.

Competing interests: The authors declare that they have no competing interests.

Funding: This study was supported in part by National Natural Science Foundation of China (81802255), Young Talents in Shanghai (2019QNBJ), 'Dream Tutor' Outstanding Young Talents Program (fkyq1901), Clinical Research Project of Shanghai Pulmonary Hospital (fk18005), Key Discipline in 2019 (oncology), Project of Shanghai Municipal Science and Technology Commission (Project of Municipal Science and Technology Commission), and Scientific research project of Shanghai Pulmonary Hospital (fkcx1903).

Authors' contributions: YC, XT, and YH conceived the concepts of the research. YH, HG, and LD designed the research. HQ and CD established PDX models. HG performed the image preprocessing. HG and LD contributed to the machine learning. All authors analyzed the data and wrote the paper.

Acknowledgements: Not applicable.

References

1. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. **Cancer Statistics, 2007.** *CA: A Cancer Journal for Clinicians* 2007, 57:43–66.
2. Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. Translational lung cancer research. 2016;5:288–300.
3. Politi K, Herbst RS. Lung cancer in the era of precision medicine. Clin Cancer Res. 2015;21:2213–20.
4. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, Gottfried M, Peled N, Tafreshi A, Cuffe S, et al. Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater. J Clin Oncol. 2019;37:537–46.
5. Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, Tsurutani J, Seto T, Satouchi M, Tada H, Hirashima T, et al. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. Lancet Oncol. 2010;11:121–8.
6. Rosell R, Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, Palmero R, Garcia-Gomez R, Pallares C, Sanchez JM, et al. Erlotinib versus standard chemotherapy as first-line treatment for

European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. Lancet Oncol. 2012;13:239–46.

7. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448:561–6.
8. Zhao ZR, Wang JF, Lin YB, Wang F, Fu S, Zhang SL, Su XD, Jiang L, Zhang YG, Shao JY, Long H. Mutation abundance affects the efficacy of EGFR tyrosine kinase inhibitor readministration in non-small-cell lung cancer with acquired resistance. Med Oncol. 2014;31:810.
9. Lim Z-F, Ma PC. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. J Hematol Oncol. 2019;12:134.
10. Liang W, Guo M, Pan Z, Cai X, Li C, Zhao Y, Liang H, Yang H, Wang Z, Chen W, et al. Association between certain non-small cell lung cancer driver mutations and predictive markers for chemotherapy or programmed death-ligand 1 inhibition. Cancer Sci. 2019;110:2014–21.
11. Cutz JC, Guan J, Bayani J, Yoshimoto M, Xue H, Sutcliffe M, English J, Flint J, LeRiche J, Yee J, et al. Establishment in severe combined immunodeficiency mice of subrenal capsule xenografts and transplantable tumor lines from a variety of primary human lung cancers: potential models for studying tumor progression-related changes. Clin Cancer Res. 2006;12:4043–54.
12. Cassidy JW, Caldas C, Bruna A. Maintaining Tumor Heterogeneity in Patient-Derived Tumor Xenografts. Cancer Res. 2015;75:2963–8.
13. Chijiwa T, Kawai K, Noguchi A, Sato H, Hayashi A, Cho H, Shiozawa M, Kishida T, Morinaga S, Yokose T, et al. Establishment of patient-derived cancer xenografts in immunodeficient NOG mice. Int J Oncol. 2015;47:61–70.
14. Ny L, Rizzo LY, Belgrano V, Karlsson J, Jespersen H, Carstam L, Bagge RO, Nilsson LM, Nilsson JA. Supporting clinical decision making in advanced melanoma by preclinical testing in personalized immune-humanized xenograft mouse models. Ann Oncol. 2020;31:266–73.
15. Jespersen H, Lindberg MF, Donia M, Söderberg EMV, Andersen R, Keller U, Ny L, Svane IM, Nilsson LM, Nilsson JA. Clinical responses to adoptive T-cell transfer can be modeled in an autologous immune-humanized mouse model. Nat Commun. 2017;8:707.
16. Fichtner I, Rolff J, Soong R, Hoffmann J, Hammer S, Sommer A, Becker M, Merk J. Establishment of patient-derived non-small cell lung cancer xenografts as models for the identification of predictive biomarkers. Clin Cancer Res. 2008;14:6456–68.
17. John T, Kohler D, Pintilie M, Yanagawa N, Pham NA, Li M, Panchal D, Hui F, Meng F, Shepherd FA, Tsao MS. The ability to form primary tumor xenografts is predictive of increased risk of disease recurrence in early-stage non-small cell lung cancer. Clin Cancer Res. 2011;17:134–41.
18. Zhang XC, Zhang J, Li M, Huang XS, Yang XN, Zhong WZ, Xie L, Zhang L, Zhou M, Gavine P, et al. Establishment of patient-derived non-small cell lung cancer xenograft models with genetic aberrations within EGFR, KRAS and FGFR1: useful tools for preclinical studies of targeted therapies. J Transl Med. 2013;11:168.

19. Izumchenko E, Paz K, Ciznadija D, Sloma I, Katz A, Vasquez-Dunddel D, Ben-Zvi I, Stebbing J, McGuire W, Harris W, et al. Patient-derived xenografts effectively capture responses to oncology therapy in a heterogeneous cohort of patients with solid tumors. Ann Oncol. 2017;28:2595–605.
20. Yu SM, Jung S-H, Chung Y-J. Comparison of the Genetic Alterations between Primary Colorectal Cancers and Their Corresponding Patient-Derived Xenograft Tissues. Genomics informatics. 2018;16:30–5.
21. Hidalgo M, Amant F, Biankin AV, Budinská E, Byrne AT, Caldas C, Clarke RB, de Jong S, Jonkers J, Mælandsmo GM, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer discovery. 2014;4:998–1013.
22. Park B, Jeong BC, Choi YL, Kwon GY, Lim JE, Seo SI, Jeon SS, Lee HM, Choi HY, Lee KS. Development and characterization of a bladder cancer xenograft model using patient-derived tumor tissue. Cancer Sci. 2013;104:631–8.
23. Tomaschek F, Hendrix P, Baayen RH. Strategies for addressing collinearity in multivariate linguistic data. Journal of Phonetics. 2018;71:249–67.
24. Sain H, Purnami SW. Combine Sampling Support Vector Machine for Imbalanced Data Classification. Procedia Computer Science. 2015;72:59–66.
25. James D. Malley KGM, Sinisa Pajevic: *Statistical Learning for Biomedical Data.* Cambridge University Press; 2011.
26. McFadden DG, Papagiannakopoulos T, Taylor-Weiner A, Stewart C, Carter SL, Cibulskis K, Bhutkar A, McKenna A, Dooley A, Vernon A, et al. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. Cell. 2014;156:1298–311.
27. Fu S, Zhao J, Bai H, Duan J, Wang Z, An T, Wang J. High-fidelity of non-small cell lung cancer xenograft models derived from bronchoscopy-guided biopsies. Thorac Cancer. 2016;7:100–10.
28. **Immune Microenvironment Differences Between Squamous and Non-squamous Non–small-cell Lung Cancer and Their Influence on the Prognosis.** *Clinical Lung Cancer* 2019, **20**:48–58.
29. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, Shafi S, Johnson DH, Mitter R, Rosenthal R, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med. 2017;376:2109–21.
30. Milholland B, Auton A, Suh Y, Vijg J. Age-related somatic mutations in the cancer genome. Oncotarget. 2015;6:24627–35.
31. Fane M, Weeraratna AT. How the ageing microenvironment influences tumour progression. Nat Rev Cancer. 2020;20:89–106.
32. Mattar M, McCarthy CR, Kulick AR, Qeriqi B, Guzman S, de Stanchina E. Establishing and Maintaining an Extensive Library of Patient-Derived Xenograft Models. Frontiers in oncology. 2018;8:19–9.
33. Raynaud F, Mina M, Tavernari D, Ciriello G. Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. PLOS Genetics. 2018;14:e1007669.

34. Lin A, Wei T, Meng H, Luo P, Zhang J. Role of the dynamic tumor microenvironment in controversies regarding immune checkpoint inhibitors for the treatment of non-small cell lung cancer (NSCLC) with EGFR mutations. Mol Cancer. 2019;18:139–9.
35. Novosiadly R, Schaer D, Lu Z, Amaladas N, Luo S, Capen A, Meyer C, Manro J, Donoho G, Doman T, et al. P3.07-006 Pemetrexed Exerts Intratumor Immunomodulatory Effects and Enhances Efficacy of Immune Checkpoint Blockade in MC38 Syngeneic Mouse Tumor Model. Journal of Thoracic Oncology. 2017;12:2300.
36. Jia Y, Li X, Jiang T, Zhao S, Zhao C, Zhang L, Liu X, Shi J, Qiao M, Luo J, et al. EGFR-targeted therapy alters the tumor microenvironment in EGFR-driven lung tumors: Implications for combination therapies. Int J Cancer. 2019;145:1432–44.
37. Wang F, Diao X-Y, Zhang X, Shao Q, Feng Y-F, An X, Wang H-Y. Identification of genetic alterations associated with primary resistance to EGFR-TKIs in advanced non-small-cell lung cancer patients with EGFR sensitive mutations. Cancer Commun. 2019;39:7.
38. Musa AB. Comparative study on classification performance between support vector machine and logistic regression. Int J Mach Learn Cybernet. 2013;4:13–24.
39. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Machine Learning for Predictive Modelling based on Small Data in Biomedical Engineering. IFAC-PapersOnLine. 2015;48:469–74.
40. Cohen ME, Hudson DL: **New chaotic methods for biomedical signal analysis**. In *Proceedings* 2000 *IEEE EMBS International Conference on Information Technology Applications in Biomedicine ITAB-ITIS 2000 Joint Meeting Third IEEE EMBS International Conference on Information Technol*; *9–10 Nov. 2000*. 2000: 123–128.

Figures

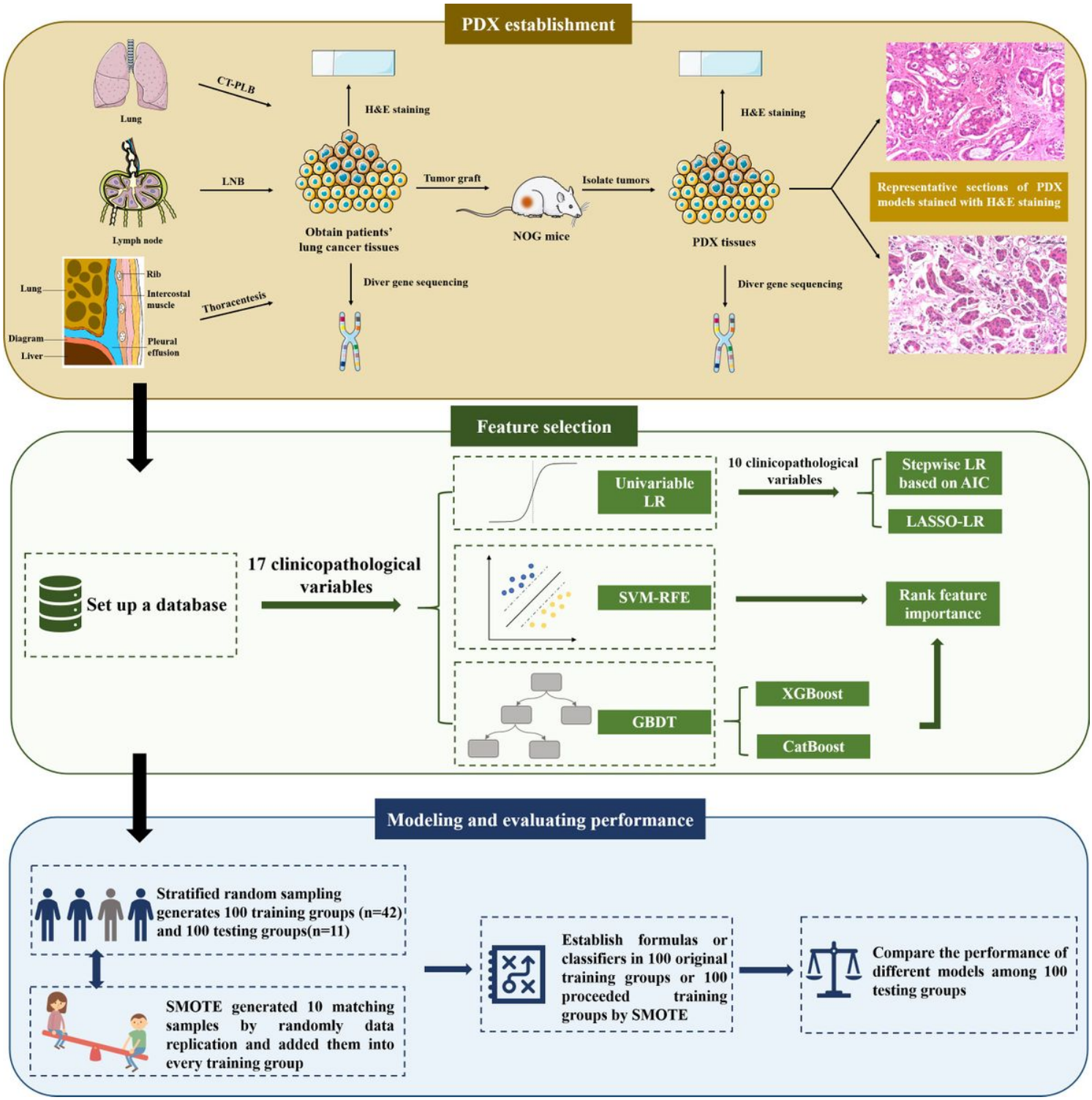

Figure 1

The study design and protocol for establishing lung cancer NOG/PDX models and machine learning. In this study, we initially obtained lung cancer tissues via percutaneous lung biopsy (CT-PLB), lymph node biopsy (LNB) or thoracentesis, and then implanted all tissues into NOG mice. After successfully establishing 53 NOG/ patient-derived tumor xenografts (PDX) models, we took all PDX tissues and then performed haemotoxylin and eosin (H&E) staining and gene sequencing to confirm whether the genotypes of PDX models matched that of patients' tumors. Further, we input 17 clinicopathological

features of patients to three machine learning methods, logistic regression (LR), support vector machine (SVM), and gradient boosting decision tree (GBDT). Afterward, we performed five algorithms of these three models, stepwise logistic regression based on the lowest Akaike information criterion (AIC), least absolute shrinkage and selection operator (LASSO)-logistic regression, support vector machine recursive feature elimination (SVM-RFE), eXtreme Gradient Boosting (XGBoost) and Gradient Boosting & Categorical Features (CatBoost), to select or rank features among all 53 samples. Next, we generated 100 training groups and 100 testing groups via stratified random sampling. Besides, we exerted synthetic minority over-sampling technique (SMOTE) to generate ten more positive class whose genotypes were different from that of parental tumors and added them into every training group. Finally, we compared the overall performance of each algorithm after trained in corresponding training groups. Note: Dr. Guo, Haoyue created this figure, and she approved it to be published in this paper.

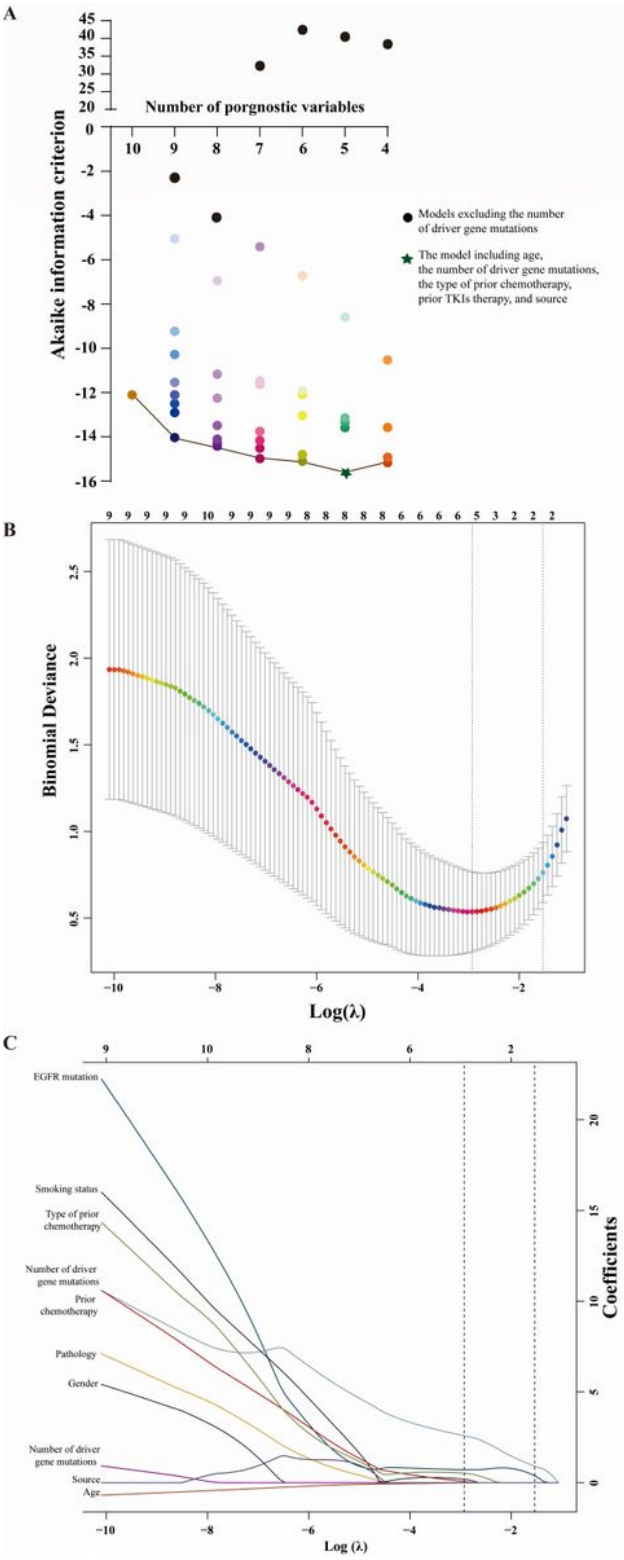

Figure 2

The feature selection based on the lowest Akaike information criterion (AIC), least absolute shrinkage and selection operator (LASSO). (A) AIC for all possible models in the stepwise multivariable logistic regression. A lower AIC indicates a better fit. Results are presented in columns defined by the number of variables in the model. Generally, the model excluding the number of driver gene mutations achieved the worst AIC, while the model containing age, the number of driver gene mutations, the type of prior

chemotherapy, prior tyrosine kinase inhibitors (TKIs) therapy, and the source was the one with lowest AIC among all potential models. (B) Factor selection using the LASSO logistic regression model. λ was the optimal penalization coefficient. The binomial deviance was plotted versus log (λ). The dotted vertical lines were plotted at the optimal λ values based on the minimum criteria and one standard error of the minimum criteria. The left vertical line represents the minimum error, and the right vertical line represents the cross-validated error within one standard error of the minimum. The optimal λ value of 1 standard error of the minimum was selected. (C) LASSO coefficients of 17 candidate variables. The right dotted vertical line was plotted at one standard error of the minimum, resulting in 2 nonzero coefficients, the number of driver gene mutations and epidermal growth factor receptor (EGFR) gene mutations. Note: Dr. Guo, Haoyue created this figure, and she approved it to be published in this paper.

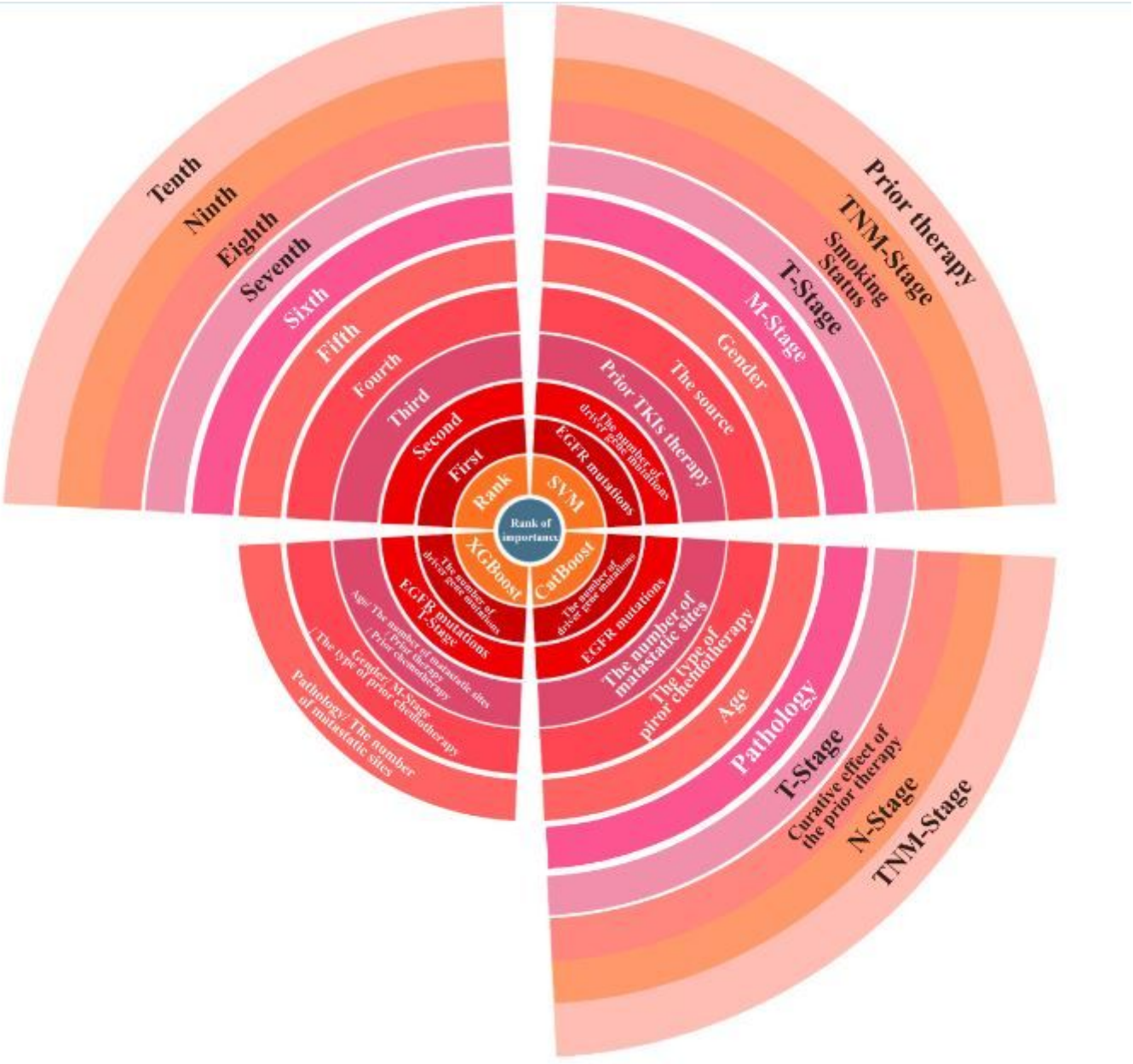

Figure 3

The rank of importance top ten variables based on support vector machine (SVM), eXtreme Gradient Boosting (XGBoost), and Gradient Boosting & Categorical Features (CatBoost). The chart showed the most critical ten variables according to the above three algorithms, where the same color represented the same ranking. Note: Dr. Guo, Haoyue created this figure, and she approved it to be published in this paper.

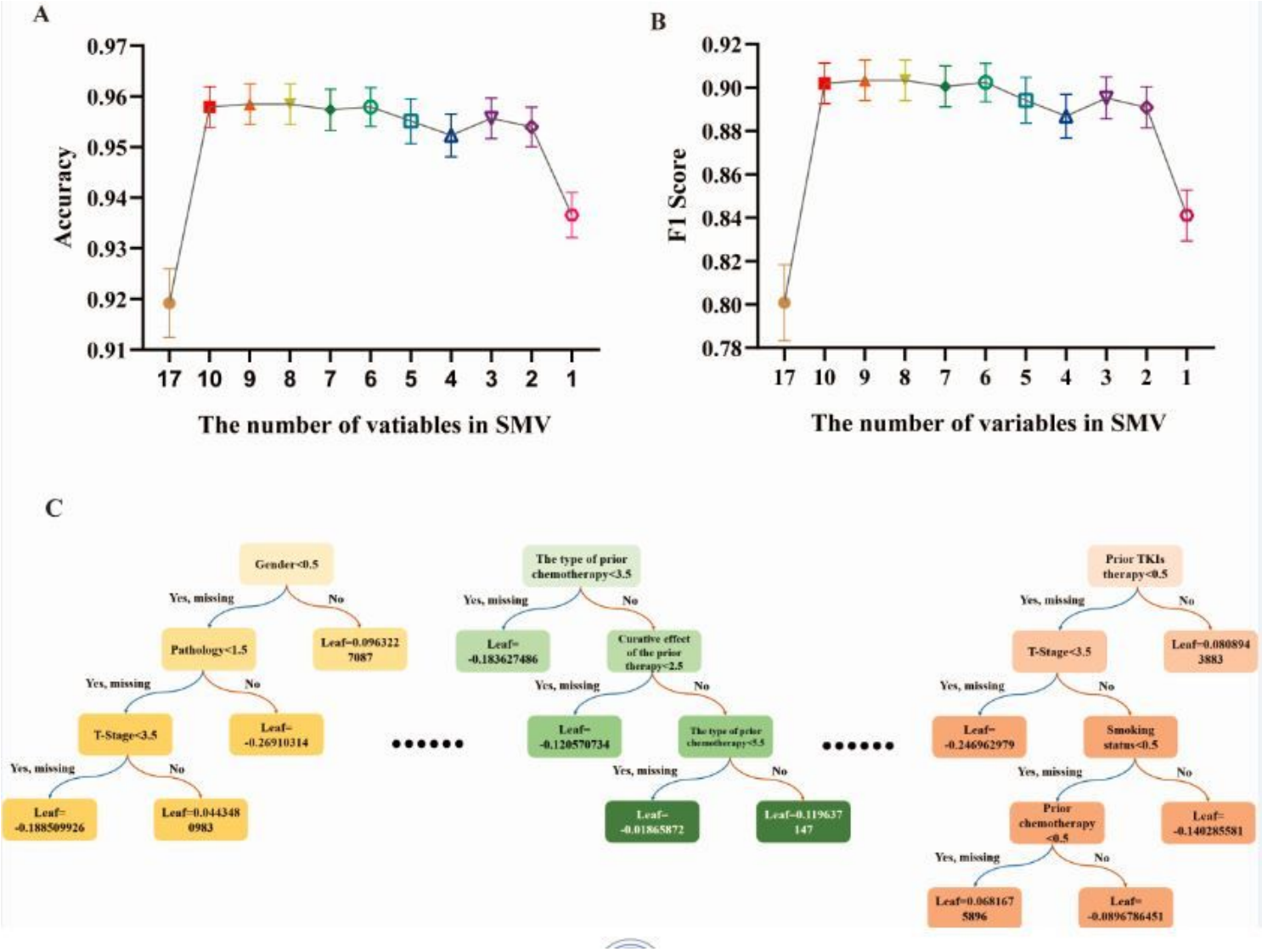

Figure 4

The modeling precess of support vector machine (SVM), and gradient boosting decision tree (GBDT). (A) The mean predictive accuracy of SMV based on the different number of variables in 100 testing groups. The 8-feature SMV was presented with the highest accuracy with the least variables. (B) The mean F1 score of SMV based on the different number of variables in 100 testing groups. The 8-feature SMV was presented with the highest F1 score with the least variables. (C) The figure showed 3 out of 100 classification and regression trees (CARTs) obtained by eXtreme gradient Boosting (XGBoost) model training. After inputting the test sample into each CART tree, we can get the predicted score of each sample at the leaf node. After weighing the total score of 100 trees, we can get the overall score of each sample and the corresponding classification. Note: Dr. Guo, Haoyue created this figure, and she approved it to be published in this paper.

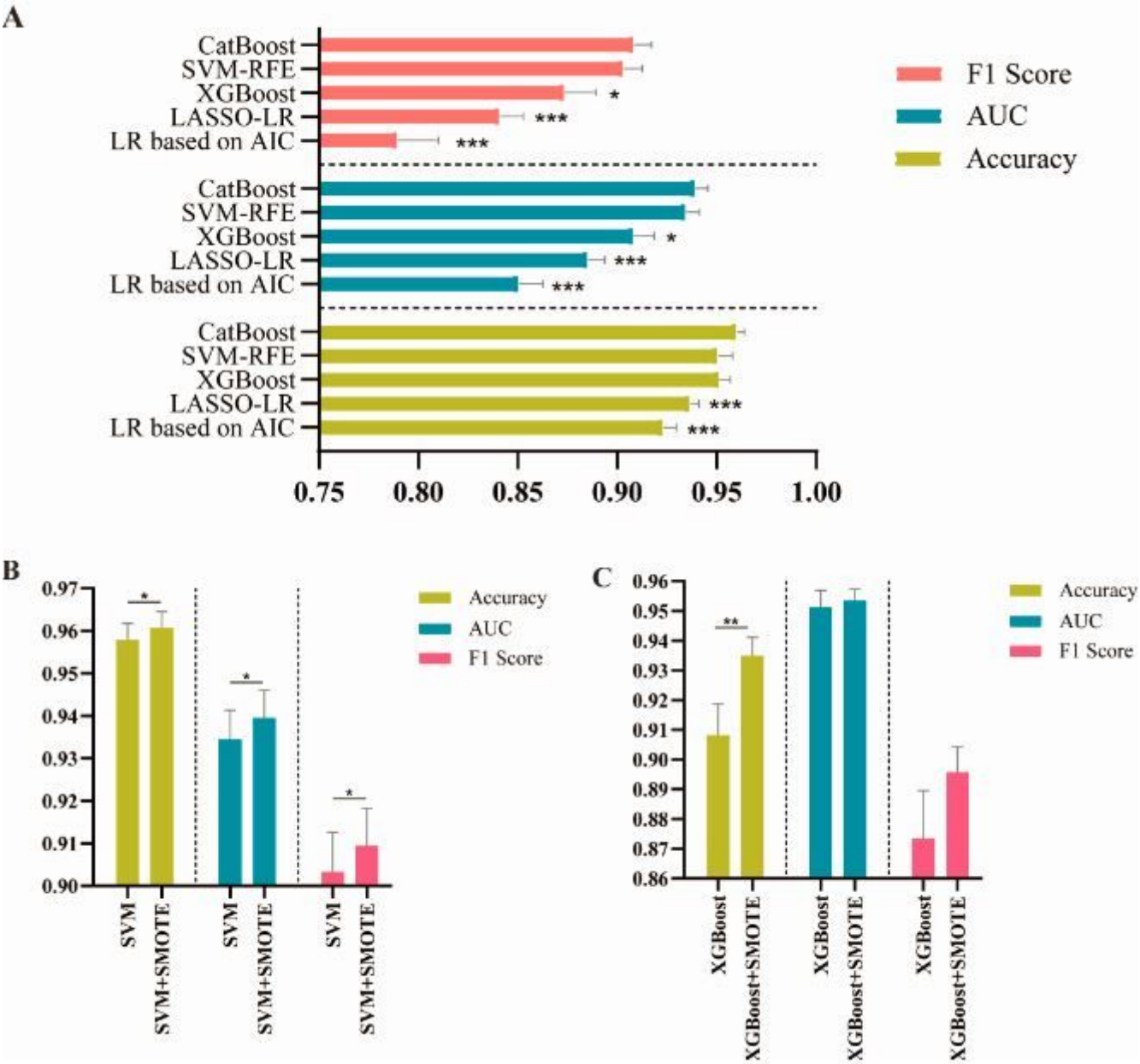

Figure 5

The comparison of performance among different models. (A) According to the predictive accuracy, the area under the receiver operating characteristic curve (AUC), and F1 score in 100 testing groups, Gradient Boosting & Categorical Features (CatBoost) and 8-feature support vector machine (SVM) exhibited the better performance than eXtreme Gradient Boosting (XGBoost), least absolute shrinkage and selection operator-logistic regression (LASSO-LR), and logistic regression (LR) based on the lowest Akaike information criterion (AIC). *: P<0.05, **: P<0.01, ***: P<0.001, based on the paired-sample t-test between CatBoost and other models. (B) With the introduction of SMOTE, the accuracy, AUC, and F1 score of 8-feature SVM in 100 testing groups acquired a significant improvement. *: P<0.05, **: P<0.01, ***: P<0.001, based on the paired-sample t-test. (C) With the introduction of synthetic minority over-sampling technique (SMOTE), the AUC of 8-feature XGBoost in 100 testing groups acquired a significant improvement. *:

P<0.05, **: P<0.01, ***: P<0.001, based on the paired-sample t-test. Note: Dr. Guo, Haoyue created this figure, and she approved it to be published in this paper.

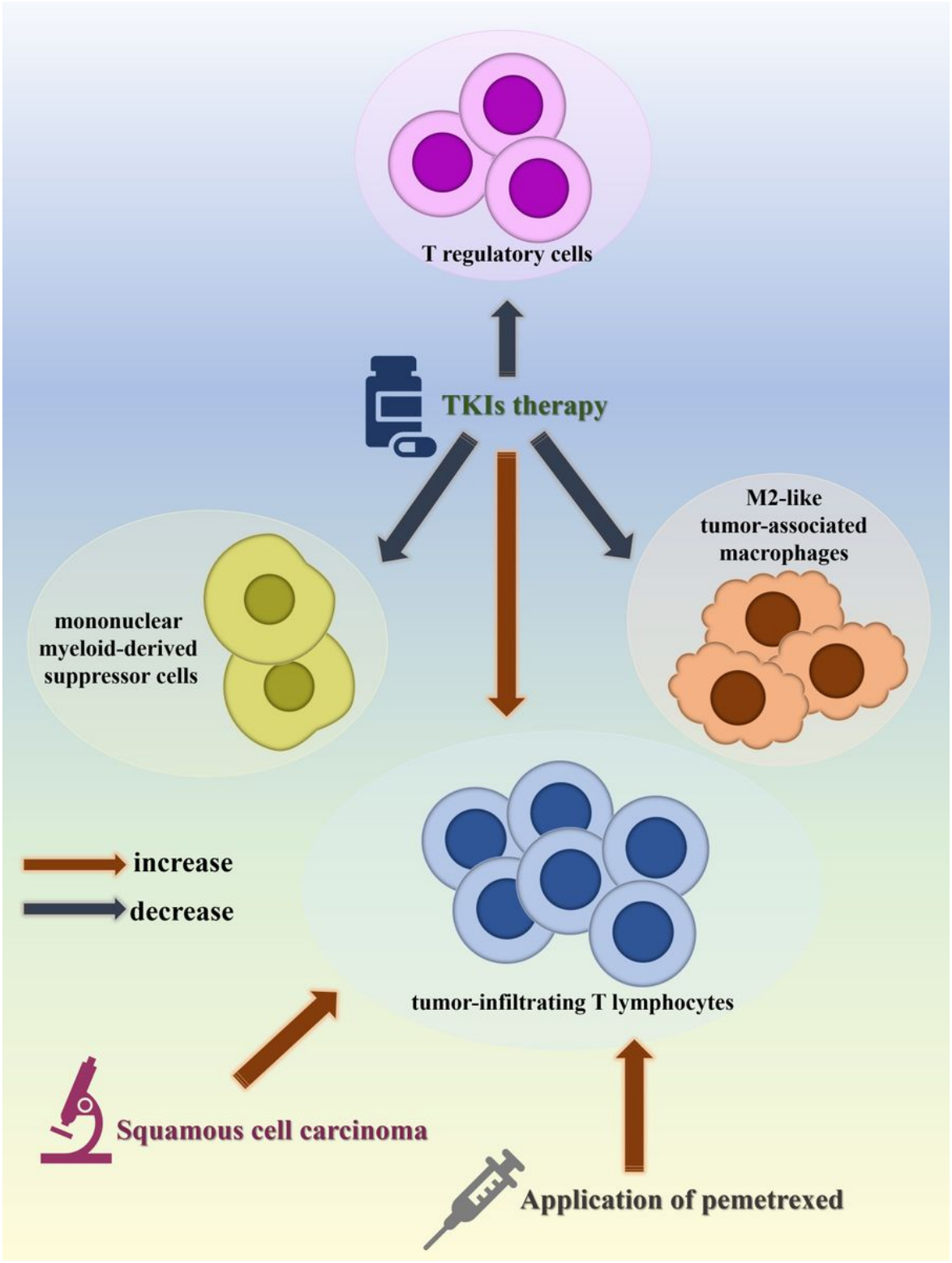

Figure 6

The association between the factor contributing to the inconsistency of driver gene mutation and the tumor microenvironment (TME). According to the univariable and multivariable logistic regression, squamous cell carcinoma, application of pemetrexed, and prior tyrosine kinase inhibitors therapy were

risk factors for the non-matching genotypes between patient-derived tumor xenografts (PDX) models and parental tumors. All of the above three factors raised tumor-infiltrating T lymphocytes. Moreover, tyrosine kinase inhibitors (TKIs) could also decrease FoxP3+ T regulatory cells (Tregs), mononuclear myeloid-derived suppressor cells (CD11b+Ly6-G-Ly6Chigh), and M2-like tumor-associated macrophages (CD206+) in the TME. Note: Dr. Guo, Haoyue created this figure, and she approved it to be published in this paper.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx