Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
New Technology of Tumor Diagnosis and Treatment—Article

基于人工智能的肺癌NOG/PDX模型驱动基因匹配预测

何雅億[a,b], 郭皓越[a,b], 刁丽[c], 陈宇[d], 朱俊杰[e], Hiran C. Fernando[g], Diego Gonzalez Rivas[e,h], 祁辉[i], 戴春雷[i], 汤旭蓁[i], 朱军[a,b], 戴家威[j], 何侃[j], Dan Chan[k], 杨洋[e,f,*]

[a] School of Medicine, Tongji University, Shanghai 200092, China
[b] Department of Medical Oncology, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai 200433, China
[c] Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[d] Spine Center, Orthopedic Department, Shanghai Changzheng Hospital, Shanghai 200003, China
[e] Department of Thoracic Surgery, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai 200433, China
[f] School of Materials Science and Engineering, Tongji University, Shanghai 201804, China
[g] Department of Thoracic Surgery, Allegheny General Hospital, Pittsburgh, PA 15212, USA
[h] Department of Thoracic Surgery and Minimally Invasive Thoracic Surgery Unit (UCTMI), Coruña University Hospital, Coruña 15006, Spain
[i] Oncology and Immunology BU, Research Service Division, WuXi Apptec, Shanghai 200131, China
[j] SJTU–Yale Joint Center for Biostatistics and Data Science, Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[k] Division of Medical Oncology, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

ARTICLE INFO

Article history:
Received 8 December 2020
Revised 5 May 2021
Accepted 20 June 2021
Available online 18 August 2021

关键词
机器学习
患者源性肿瘤异种移植物
NOG小鼠

摘要

患者源性肿瘤异种移植物（PDX）是癌症药物发现和筛查的有力工具。然而，目前的研究对PDX的基因型错配知之甚少，导致PDX使用过程中产生巨大的经济损失。在此，本研究建立了53例肺癌患者的PDX模型，基因型匹配率为79.2%（42/53）。此外，检查了17个临床病理学特征，并基于最低赤池信息量准则（AIC）、最小绝对收缩和选择算子（LASSO）-逻辑回归（LR）、支持向量机（SVM）递归特征消除（SVM-RFE）、极端梯度增强（XGBoost）、梯度增强和分类特征（CatBoost），以及合成少数过采样技术（SMOTE）输入逐步逻辑回归模型。最后，通过100个试验组的准确度、受试者工作特征曲线下面积（AUC）和F1评分评价所有模型的性能。两个多变量LR模型显示，年龄、驱动基因突变的数量、表皮生长因子受体（*EGFR*）基因突变、既往化疗的类型、既往酪氨酸激酶抑制剂（TKI）治疗和样本来源是强有力的预测因素。此外，CatBoost（平均精度= 0.960；平均AUC = 0.939；平均F1分数= 0.908）和八特征SVM-RFE（平均精度= 0.950；平均AUC = 0.934；平均F1分数= 0.903）在算法中表现出最好的性能。同时，除CatBoost外，SMOTE的应用提高了大多数模型的预测能力。基于SMOTE，单一模型的集成分类器达到了最高的准确度（平均值= 0.975）、AUC（平均值= 0.949）和F1评分（平均值= 0.938）。总之，本文建立了一个最佳预测模型来筛选肺癌患者的NOD/Shi-scid白细胞介素-2受体（IL-2R）γ^{null}（NOG）/PDX模型，并为建立预测模型提供了一种通用方法。

1. 引言

肺癌是人类癌症死亡的原因之一，全球每年有超过100万人死于肺癌[1]。大约85%的肺癌为非小细胞肺癌（NSCLC），小细胞肺癌（SCLC）占肺癌的15% [2]。最近，随着驱动基因的引入和分子检测技术的进步，肺癌患

* Corresponding author.
E-mail address: timyangsh@tongji.edu.cn (Y. Yang).

英文原文：Engineering 2022, 15(8): 102–114
引用本文：Yayi He, Haoyue Guo, Li Diao, Yu Chen, Junjie Zhu, Hiran C. Fernando, Diego Gonzalez Rivas, Hui Qi, Chunlei Dai, Xuzhen Tang, Jun Zhu, Jiawei Dai, Kan He, Dan Chan, Yang Yang. Prediction of Driver Gene Matching in Lung Cancer NOG/PDX Models Based on Artificial Intelligence. *Engineering*, https://doi.org/10.1016/j.eng.2021.06.017

者基于化疗的治疗模式发生了转变[3–4]，尤其是对于表皮生长因子受体（*EGFR*）基因突变[5–6]和间变性大细胞淋巴瘤激酶（ALK）基因重排的患者[7]。然而，靶向治疗面临一系列困难，包括不同的个体反应和频繁的获得性耐药[8–9]。免疫检查点抑制剂（ICI）被推荐用于肺癌患者[10]。然而，只有大约20%的肺癌患者对免疫治疗有反应。确定肺癌患者的耐药机制具有重要意义。

临床前动物模型对于药物筛选至关重要。患者源性肿瘤异种移植物（PDX）已成为一种准确的临床前系统，能够维持亲代肿瘤的分子、遗传和组织病理学异质性[11–12]。而且，新一代的超免疫缺陷小鼠[称为NOD/Shi-scid白细胞介素-2受体（IL-2R）γ^{null}（NOG）/PDX小鼠]的特征是白细胞介素-2受体共同γ链和多种免疫细胞[如T细胞、B细胞、自然杀伤（NK）细胞、巨噬细胞、树突状细胞]的缺失，被认为是构建癌症免疫治疗PDX模型的极好选择[13]。与裸鼠和传统的重症联合免疫缺陷（SCID）小鼠相比，NOG小鼠在ICI、过继性T细胞治疗（ACT）和其他免疫治疗的研究中表现出突出的潜力，因为肿瘤浸润T淋巴细胞（TIL）可以在异种移植物形成后被连续移植到其中[14–15]。

尽管如此，PDX建立的低成功率（20%~40%）[16–18] 和已建立的PDX与原始肿瘤之间驱动基因突变不一致的显著发生率（10%~20%）[19–20]是一个研究尚未充分的问题。由于生成PDX模型的方案耗时、耗力，且成本高昂[21]，因此驱动基因突变的不一致性对于研究人员、医生和患者来说是一个棘手的问题。然而，多种因素，包括性别、吸烟史、病理学、肿瘤-淋巴结-转移（TNM）分期、肿瘤分级、肿瘤样本质量和*EGFR*基因突变，已被证明与肿瘤成功植入率相关[17–18,22]。对于这些因素是否有助于PDX模型（尤其在NOG小鼠中建立的模型）中驱动基因突变的一致性的问题尚未得到验证。本研究使用机器学习（ML）算法，包括多变量逻辑回归（LR）、支持向量机（SVM）递归特征消除（SVM-RFE）、梯度增强决策树（GBDT）和合成少数过采样技术（SMOTE），建立了预测NOG/PDX模型与患者肿瘤之间驱动基因突变不一致的有力工具（图1）。

2. 材料与方法

2.1. 患者样本

53例患者的肺癌组织或细胞通过2018年8月至2019年10月在上海市肺科医院（中国）进行的计算机断层扫描（CT）引导下经皮肺活检（CT-PLB）、淋巴结活检（LNB）或胸腔穿刺术获取。所有患者均提供了书面知情同意书，授权采集和使用其组织用于研究。本研究获得了上海市肺科医院伦理/许可委员会的批准（批准文号：NO K18-203）。此外，本研究按照1964年《赫尔辛基宣言》的伦理标准进行。

2.2. 组织样品的制备

将经支气管活检（TBB）、CT-PLB和LNB收获的组织分为三块。将第一块切成50~100 mm^3的碎片，浸入冷冻的Bambanker培养基（产品目录号：BBH01；Nippon Genetics, Japan），然后保存在液氮中直至植入免疫缺陷小鼠体内。对第二块进行立即液氮冷冻，用于DNA/RNA提取。将第三块用于生成福尔马林固定石蜡包埋（FFPE）切片进行病理学评估。

2.3. 恶性胸腔积液的制备

恶性胸腔积液（MPE）的制备和培养按照如前所述的方法进行。通过胸腔穿刺术抽取200~1000 mL胸腔积液。以3000 $r\cdot min^{-1}$的速度离心样品10 min，然后重悬于磷酸盐缓冲盐水（PBS）中。使用Ficoll-Paque PLUS（GE Healthcare Bio-Sciences, Sweden）通过密度梯度离心法从样本间期层分离肿瘤细胞。用PBS清洗后，在含10%胎牛血清（FBS; Thermo Fisher Scientific, USA）和10 $ng\cdot mL^{-1}$表皮生长因子（EGF）的Roswell Park Memorial Institute（RPMI）-1640中培养肿瘤细胞，密度为$1\times10^6\sim2\times10^6$个细胞/板。

2.4. NOG/PDX建立

本研究中的所有动物实验均遵循机构动物护理和使用委员会（IACUC）的指导原则。在6~8周龄雌性NOG小鼠（Charles River, China）中建立PDX模型。将冷冻组织在37 ℃下解冻，并直接皮下植入NOG小鼠的无菌皮肤中（每个肿瘤样本$n = 4\sim5$）。同时，在PBS中清洗从MPE中分离的肿瘤细胞一次，然后注射（5106个细胞）至每只NOG小鼠的右腹侧（每份MPE样本$n = 4\sim5$）。

将初始肿瘤植入的NOG小鼠维持120天，每周测量一次肿瘤大小。使用以下公式计算肿瘤体积（TV）：TV =（长度×宽度的二次方）/2（长度为最长直径，而宽度为最短直径）。当肿瘤大小达到700~800 mm^3时，对异种移植肿瘤进行传代，本研究中使用的PDX模型从第3代（P3）传至第5代（P5）。将每次传代的PDX肿瘤分离成三块。将第一块植入另一只NOG小鼠体内进行传代。将第二块进行立即液氮冷冻，用于DNA/RNA提取。将第三块用于生成FFPE切片进行病理学评估。

所有动物护理和实验均按照上海市肺科医院伦理/许

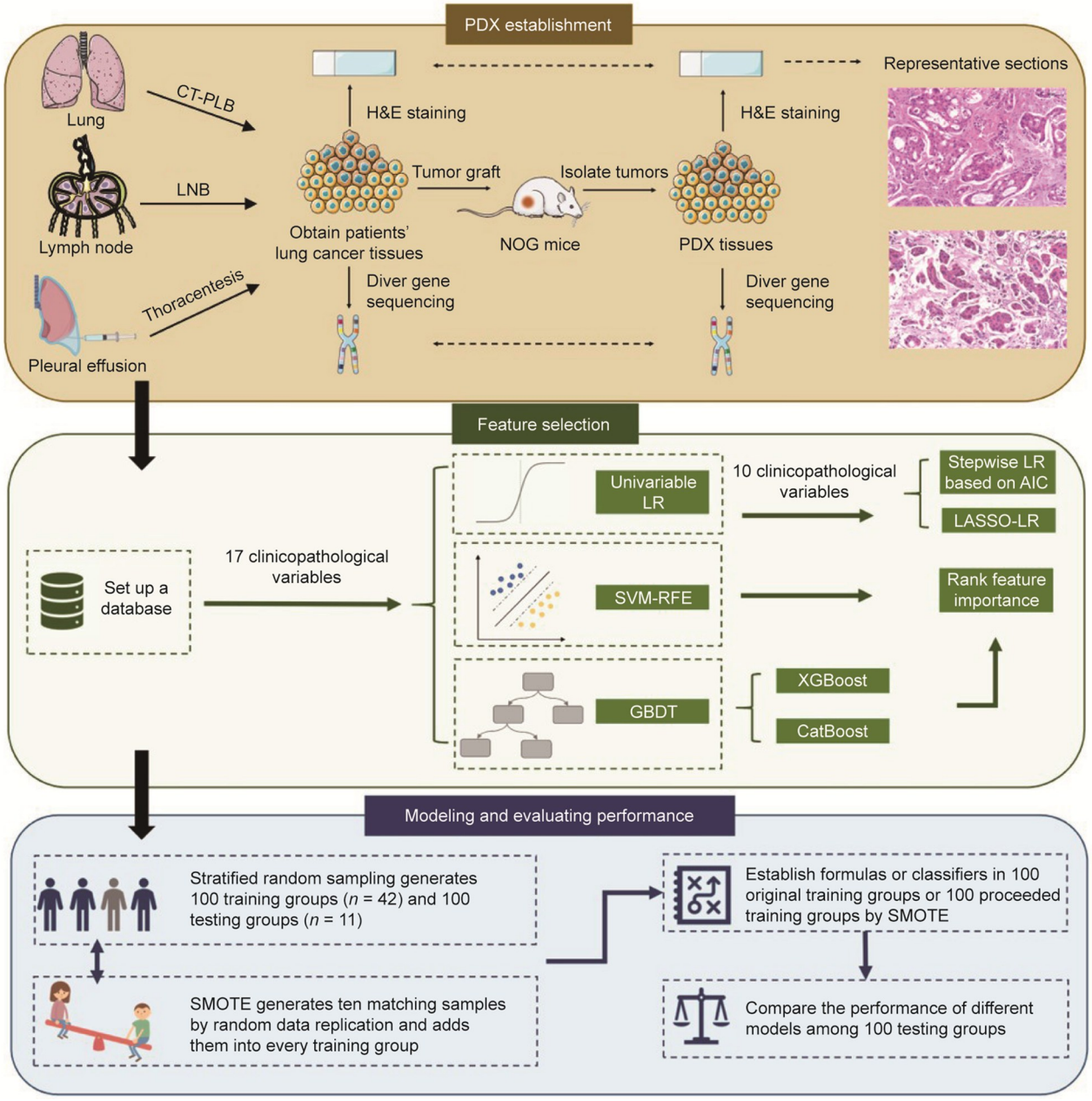


图1. 建立肺癌NOG/PDX模型和ML的研究设计及方案。本研究最初通过计算机断层扫描（CT）引导的经皮肺活检（CT-PLB）、淋巴结活检（LNB）或胸腔穿刺术获得肺癌组织，然后将所有组织植入NOG小鼠体内。成功建立53个NOG/PDX模型后，取所有PDX组织，然后进行苏木精伊红（H&E）染色及基因测序，证实PDX模型的基因型是否与患者肿瘤相匹配。然后，将患者的17个临床病理特征输入三种ML方法——LR、SVM-RFE和GBDT中。之后，对这三种模型进行了5种算法，基于最低赤池信息量准则（AIC）的逐步LR、最小绝对收缩和选择算子（LASSO）-LR、SVM-RFE、极端梯度增强（XGBoost）、梯度增强和分类特征（CatBoost），在所有53个样本中选择或排序特征。接下来，通过分层随机抽样生成了100个训练组和100个测试组。还采用了SMOTE，生成了10个更多的阳性类别，其基因型与亲代肿瘤不同，并将其添加到每个训练组中。最后，比较了相应训练组训练后的各算法的整体性能。

可委员会批准的动物方案进行。

2.5. DNA和RNA提取

对肺癌组织和PDX组织进行病理复查，确保肿瘤细胞占肿瘤的80%以上，DNA提取前未发生明显的肿瘤坏死。使用QIAamp DNA迷你试剂盒（Qiagen-51306, Germany）从每份组织样本中提取基因组DNA。使用Nanodrop ND-1000 UV/VIS分光光度计（Thermo Scientific, USA）测定DNA样品的数量和纯度。使用1%琼脂糖凝胶电泳确认DNA片段的完整性。将DNA浓度归一化至20 ng·μL^{-1}，并在20 ℃下储存直至使用。通过扩增难治性突变系统（ARMS）和突变富集液相芯片聚合酶链反应（PCR）筛选*EGFR*（外显子18、19、20和21）和*ALK*融合（*EML4-ALK*）的“热点”（hot spot）突变。

2.6. 基因型匹配和错配的定义

本研究将基因型匹配定义为PDX模型中*EGFR*和*ALK*的突变类型与相应患者完全相同。定义了三种基因型错配：①患者的原始驱动基因突变在PDX模型中不存在；②在PDX模型中检测到驱动基因突变，而在相应患者中未检测到；③驱动基因突变在PDX模型和相应患者中均出现，但PDX模型与患者驱动基因的数量和（或）类型不一致。

2.7. 机器学习

2.7.1. 基于最低赤池信息量准则的逐步LR

赤池信息量准则（AIC）是评分和选择模型的指标，使用以下公式计算：AIC = −2/N ×（log-likelihood）+ 2K/N。式中，N是示例数量；K是模型中变量的数量加上截距；对数似然（log-likelihood）是模型拟合的指标，通常从统计输出中获得。使用R软件的“MASS”包根据最低AIC进行逐步LR。

2.7.2. 最小绝对收缩和选择算子-LR

驱动基因突变用作LR模型中的因变量Y输入，编码为0表示缺失（一致性），编码为1表示存在（不一致性）。给定协变量X_i，驱动基因突变不一致的概率计算如下：$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_kX_k$。

将逻辑最小绝对收缩和选择算子（LASSO）估计量$\beta_0, \ldots, \beta_k$定义为负对数似然的最小值：$\sum_{i=1}^{n}-y_i(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)+\lg[1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)]$受制于$\sum_{j=1}^{k}|\beta_j|\le\lambda$。

式中，$\lambda>0$是控制估计量稀疏性的调整参数（即数值为零的系数数量），实际上是通过使用验证样本或交叉验证选择的[23]。为了获得逻辑LASSO估计量，使用了R软件中的“glmnet”包。

2.7.3. 支持向量机递归特征消除

SVM-RFE是一种基于SVM的特征消除方法，通过从初始特征集中选择性能最佳的肽集，提供分类作为输出。据报道，该方法是解决生物信息学中过度拟合的最佳分类算法之一。本研究使用了用于SVM-RFE的Python软件的“sklearn. feature_selection”包（内核 =“线性”，C = 0.3）。

2.7.4. 极端梯度增强

极端梯度增强（XGBoost）用于监督学习问题；本研究使用XGBoost来分类驱动基因突变是否在PDX模型和亲代肿瘤中匹配。在本研究中，每次迭代使用树增压器，XGboost应用Python软件的“xgboost”包，参数如下：max_depth = 6, subsample = 1.0, min_child_weight = 1.0, γ = 0.3, learning_rate = 0.2, n_estimators = 100, colsample_bytree = 0.25, eval_metric =“auc”。

2.7.5. 梯度增强和分类特征

GBDT库的另一种算法，即梯度增强和分类特征（CatBoost）被用于特征选择和预测工具的建立。本研究使用了用于CatBoost的Python软件的“catboost”包。CatBoost的详细参数如下：loss_function =“Logloss”, eval_metric =“AUC”, learning_rate = 0.1, depth = 6, iterations = 100, border_count = 20, subsample = 1.0, colsample_bylevel = 1.0, random_strength = 0.7, scale_pos_weight = 1, reg_lambda = 10。

2.7.6. 合成少数过采样技术

SMOTE是一种过采样方法，该方法利用采样方法，通过随机数据复制来增加正类的数量，使正数据量等于负数据量。SMOTE算法由Chawla等[24]首次提出。这种方法通过构建综合数据-次要数据复制来工作。SMOTE算法通过定义每个正类的k最近邻，然后对正类和随机选择的k最近邻之间所需的百分比进行数据合成。一般可用公式表示为$X_{syn}=X_i+(X_{knn}-X_i)\times\delta$，其中$X_{syn}$代表合成数据点，$X_i$和$X_{knn}$是在$k$最近邻中随机挑选的原例和最近邻数据点，而$\delta$是0和1之间的随机数。本研究中，使用Python软件的“smote_variants”包为SMOTE添加了另外10个阳性样本到训练数据中。

2.7.7. 自举重新采样

在本研究中，使用Python软件的自助法，通过重新采样对100个训练组（n = 35；7个不匹配样本和28个匹配样本）和100个测试组（n = 18；4个不匹配样本和14个匹配样本）进行分层。

2.8. 模型性能评价

为了评价本研究中所有预测模型的性能，计算了相关指标（表1）。

进行的计算如下：

（1）准确度 = [真阳性计数（TP）+真阴性计数（TN）] / [TP + TN +假阳性计数（FP）+假阴性计数（FN）]

表1 本研究中性能指标的定义

Predicted class	True class	
	Positive	Negative
Positive	True positive count (TP)	False positive count (FP)
Negative	False negative count (FN)	True negative count (TN)

（2）精密度= TP / (TP + FP)

（3）召回率= TP / (TP + FN)

（4）F1 评分=（2 ×精密度×召回率）/（精密度+召回率）

（5）受试者工作特征曲线（ROC）以假阳性率为横坐标，真阳性率为纵坐标。

通过Python软件的“sklearn. metrics”计算ROC曲线下面积（AUC）。

2.9. 统计学分析

采用配对样本 *t* 检验比较试验组不同模型间的性能。本研究中的所有数据分析均使用统计产品和服务解决方案（SPSS，版本23.0；IBM SPSS, USA）、R软件（版本3.1.0；R Core Team, USA）、MATLAB（版本7.12.0；Mathworks, USA）和Python软件（版本2.7；Python Software Foundation, USA）进行。所有图均使用GraphPad Prism（版本8.0；GraphPad Software, USA）生成。统计学检验为双侧检验，$P < 0.05$ 为差异有统计学意义。

2.10. 计算机代码可用性

本研究的所有原始数据和代码可从以下网址获取：https://github.com/dddtqshmpmz/PDX。

3. 结果

3.1. NOG/PDX模型的建立

53例NSCLC患者的一般临床病理学特征见附录A中的表S1和表S2。患者中位年龄为66岁，83.0%（44/53）为男性。三例患者（5.7%）被诊断为TNM-1期，其他患者（94.3%）被诊断为TNM-3期或TNM-34期。40例患者（75.5%）被诊断为NSCLC，其中鳞状细胞癌（SCC）9例、腺癌（ADC）15例，其他NSCLC 16例。13例患者（24.5%）患SCLC。在所有样本中，10例患者（18.9%）的组织存在*EGFR*基因突变，一例患者（1.9%）存在*ALK*融合，其余42例样本（79.2%）为非突变组织。肿瘤转移的有39例（73.6%）。

通过CT-PLB获得48例样品（90.6%），通过LNB获得两例（3.8%），通过胸腔穿刺获得三例（5.7%）。有39例患者（73.6%）在采样前接受了治疗，包括化疗（$n = 29$）、酪氨酸激酶抑制剂（TKI）（$n = 8$）和免疫治疗（$n = 2$），而14例患者（26.4%）未接受任何治疗。

本研究纳入的所有PDX模型均被病理学家证实已成功建立（大小达到700~800 mm^3），苏木精和伊红（H&E）染色的PDX模型代表性切片如图1所示。总的驱动基因匹配率为84%（42/50）。

3.2. 模型1——LR

3.2.1. 单变量分析与基因型不匹配相关的因素

单变量LR分析表明，PDX模型与亲本肿瘤驱动基因突变不一致的危险因素为：女性、低龄、吸烟史、从LNB或胸腔穿刺获得、NSCLC（SCC除外）、*EGFR*突变、更多的驱动基因突变、既往未接受过化疗、既往接受过培美曲塞+卡铂化疗和既往TKI治疗（表2）。

表2 53例患者的特征和17个临床病理学变量的单变量LR，用于确定与PDX模型和亲代肿瘤之间驱动基因不一致相关的因素，通过单变量LR分析*P*值和比值比（OR）

Variable	Matching of driver genes between PDX models and parental tumors		*P*	OR (95% CI)
	Yes	No		
Mean age (year)	65.36	59.73	0.050	0.921 (0.848–1.000)
Gender				
Male	39	5	0.010	1.000
Female	3	6		15.600 (2.938–82.836)
Smoking status				
No	12	9	0.010	1.000
Yes	30	2		11.250 (2.113–59.884)
Source of the sample				
CT-PLB	40	8	0.042	1.000
LNB or thoracentesis	2	3		7.500 (1.704–52.377)
Pathology				
ADC	8	7	0.012	10.500 (1.076–102.478)
SCC	9	0		0
Other NSCLCs	13	3		2.769 (0.252–30.383)

续表

Variable	Matching of driver genes between PDX models and parental tumors		*P*	OR (95% CI)
	Yes	No		
SCLC	12	1		1.000
EGFR mutation				
No	41	2	< 0.001	1.000
Yes	1	9		184.500 (15.046–2 262.404)
Number of mutations				
0	41	1	< 0.001	1.000
1	1	7		287.000 (16.024–5 140.254)
2	0	3		6.620×10^{10}
T-stage				
1–2	12	1	0.148	1.000
3–4	30	10		4.000 (0.460–34.750)
N-stage				
0	5	1	0.546	1.000
1	4	0		0
2	12	4		1.667 (0.147–18.874)
3	21	6		1.429 (0.139–14.695)
M-stage				
0	13	1	0.173	1.000
1	29	10		0.223 (0.026–1.929)
TNM-stage				
1–2	3	0	0.230	1.000
3–4	39	11		455 646 925.900
Number of distant metastatic sites				
0	13	1	1.000	1.000
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		1.000 (0.020–50.397)
Prior therapy				
No	12	2	0.473	1.000
Yes	30	9		1.800 (0.338–9.581)
Prior chemotherapy				
No	15	9	0.005	8.100 (1.545–42.476)
Yes	27	2		1.000
Chemotherapy type				
EC	7	0	0.037	1.000
GC	7	0		1.000
Paclitaxel liposome	3	0		1.000
AC	3	1		538 491 658.700
Other chemotherapy	7	1		2 307 282 139.000
None	15	9		989 284 985.700
Prior TKI therapy				
No	39	6	0.005	1.000
Yes	5	3		10.833 (2.040–57.525)
Curative effect of prior therapy				

续表

Variable	Matching of driver genes between PDX models and parental tumors		*P*	OR (95% CI)
	Yes	No		
No therapy	12	2	0.387	1.000
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 (OR) are analyzed by univariate LR.

EC: etoposide and carboplatin; GC: gemcitabine and cisplatin; AC: pemetrexed and carboplatin; PR: partial response; PD: progressive disease; SD: stable disease; CI: confidence interval; T stage: size or direct extent of the primary tumor; N stage: degree of spread to regional lymph nodes; M stage: presence of distant metastasis; TNMstage: tumor, node, metastasis (TNM) staging classification according to the American Joint Committee on Cancer (AJCC).

3.2.2. 所有53个示例中的多变量选择

基于AIC的LR。为了平衡预测模型的性能和复杂性，通过计算AIC进行了逐步模型选择。根据单变量分析，有10个潜在的预测特征。图2（a）显示了向后逐步LR中每个步骤的AIC值，其中逐个删除10个预测特征，直至AIC不再降低。一般而言，排除驱动基因数量的模型呈现最差的AIC，表明驱动基因数量是一个重要的预测因子。此外，AIC选择的最佳多变量模型是五变量LR模型，包括年龄、驱动基因突变数量、既往化疗类型、既往TKI治疗和样本来源。

LASSO-LR。在LR中进行了LASSO正则化，以提高预测的准确度和可解释性。本文将单变量LR模型的10个显著特征输入到多变量LASSO-LR中。通过LASSO-LR结合十倍交叉验证筛选出10个特征中的两个特征，即*EGFR*突变和驱动基因的数量，其中最佳惩罚系数λ被确定为一个标准误差[图2（b）、（c）]。

3.3. 模型2——SVM-RFE

SVM-RFE从一个完整的特征集开始，根据维度长度的权重向量，消除每次迭代中分类的最不重要特征。根据特征重要性的排序，在图3（a）中首先删除了7个最不重要的变量，然后逐一删除了其余10个变量，以优化预测准确度。根据测试组的平均预测精度和F1评分，包括8个变量的SVM-RFE模型以最小的复杂性保持了最佳的性能[图3（b）、（c）]。结果表明，八特征SVM-RFE是所有SVM分类器中的最佳模型。

3.4. 模型3——GBDT

为了实现GBDT，使用两种常用的算法：XGBoost和CatBoost。大量实验表明，特征之间的多重共线性不会阻碍决策树的预测分类[25]。因此，在本研究中输入了XGBoost和CatBoost中的所有17个功能。基于XGBoost和CatBoost分类算法的特征等级如图3（a）所示。XGBoost和CatBoost生成的决策树的代表性结构如图3（d）所示。

3.5. 训练组建模和测试组性能评价

3.5.1. 不同模型之间的对比

根据100个测试组的AUC、准确度和F1评分，CatBoost（平均准确度 = 0.960；平均AUC = 0.939；平均F1评分= 0.908）和八特征SVM-RFE（平均准确度= 0.950；平均AUC = 0.934；平均F1评分= 0.903）显著优于其他三种模型：XGBoost（平均准确度= 0.951；平均AUC = 0.908；平均F1评分 = 0.873）、LASSO-LR（平均准确度= 0.937；平均AUC = 0.886；平均F1评分= 0.841）和基于AIC的LR（平均准确度= 0.923；平均AUC = 0.850；平均F1评分= 0.789）。虽然八特征SVM-RFE和XGBoost的准确度在统计学上是相等的，但CatBoost和八特征SVM-RFE的整体性能最好。此外，CatBoost和八特征SVM-RFE的平均准确度（$P = 0.103$）、AUC（$P = 0.066$）和F1评分（$P = 0.128$）没有显著差异[图4（a）]，表明有希望克服不平衡的小样本数据集的局限性。本文还评价了这些模型性能的偏倚[图4（b）]，除基于AIC的LR的F1评分（训练组和试验组之间的差异为11.6%）外，训练组和试验组之间的准确度、F1评分和AUC差异保持在8%以下。

3.5.2. 使用SMOTE改善模型性能

SMOTE是一种过采样方法，通过随机数据复制来增加正类的数量，使正类和负类具有相等的数量[26]。本研究应用SMOTE将另外10个阳性样本添加到训练组中，以便完成特征选择，建立每个模型，然后在原始100个测试组中测试模型（见附录A中的表S2、表S3）。LASSO-LR在用于SMOTE的平衡训练数据中具有两个相同的特征：*EGFR*突变和驱动基因突变的数量。相比之下，基于AIC

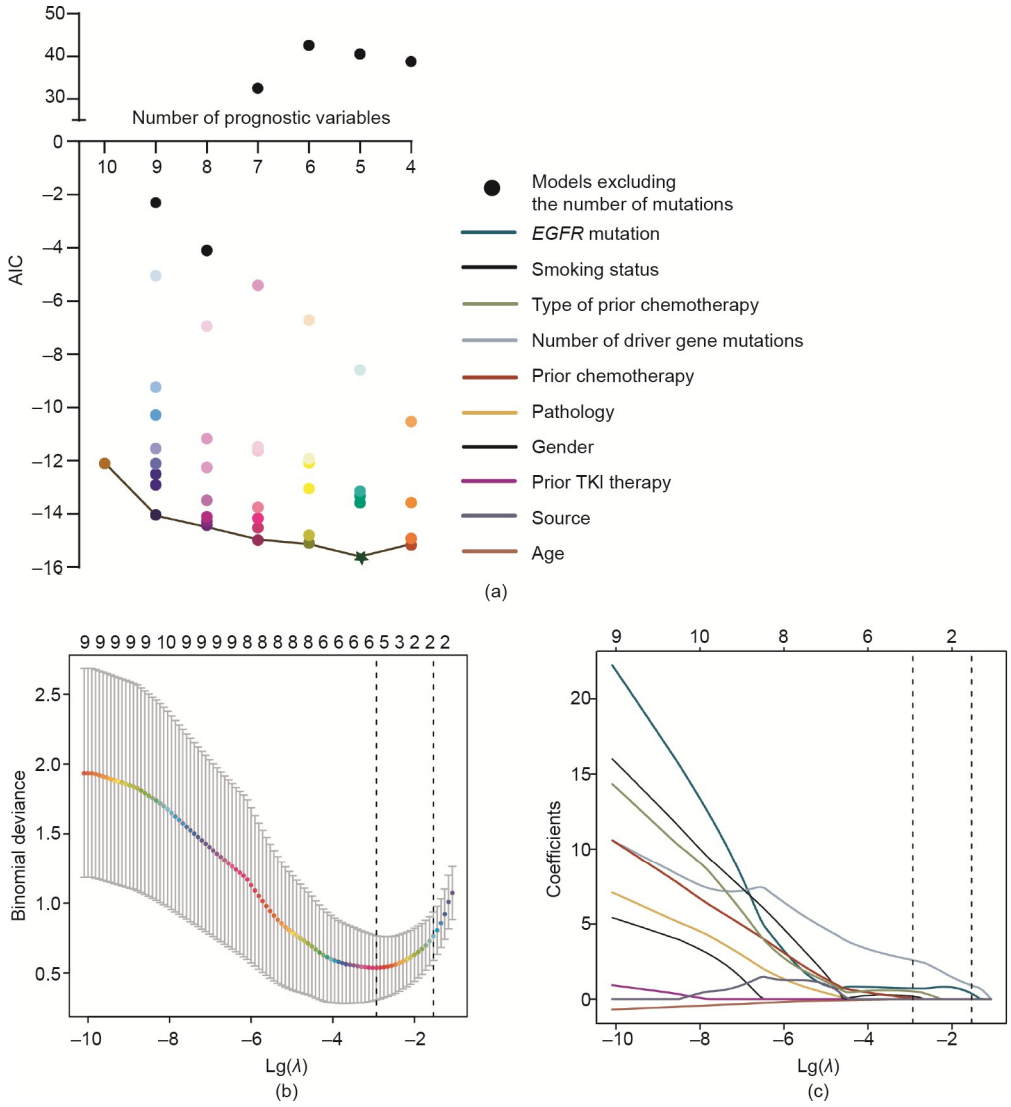


图2. 基于最低AIC和LASSO的特征选择。（a）逐步多变量LR中所有可能模型的AIC越低表示拟合越好。结果以模型中变量数量定义的列表示。一般而言，排除驱动基因突变数量的模型达到的AIC最差，而包含年龄、驱动基因突变数量、既往化疗类型、既往TKI治疗和样本来源的模型是所有潜在模型中AIC最低的模型。上横坐标是此时模型中非零系数的数量。（b）使用LASSO-LR模型的因素选择。λ为最佳惩罚系数。绘制二项式偏差相对lgλ的曲线。基于最低标准和最低标准的一个标准误差，在最佳λ值处绘制垂直虚线。左垂直线代表最小误差，右垂直线代表在最小值的一个标准误差内的交叉验证误差。选择最小值的一个标准误差的最佳λ值。上横坐标是此时模型中非零系数的数量。（c）在单变量LR中有意义的10个候选变量的LASSO系数。在最小值的一个标准误差处绘制右虚线垂直线，得到驱动基因突变数和*EGFR*基因突变数两个非零系数。

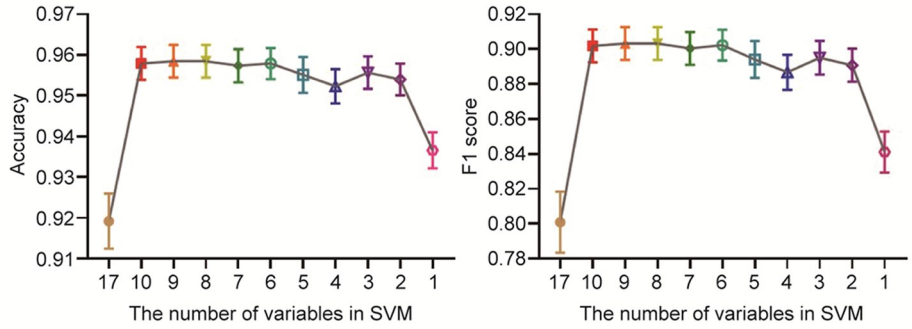
的LR从所有特征中选择了以下7个特征：性别、*EGFR*突变、驱动基因突变数量、M分期、转移部位数量、既往化疗类型、既往TKI治疗和样本来源。SVM-RFE、XGBoost和CatBoost中10个最关键的特征排名如图5所示，驱动基因突变和*EGFR*突变的数量仍然是主要的促成因素。

有趣的是，SMOTE增强了LASSO-LR（准确度：0.957 vs 0.923；AUC：0.936 vs 0.850；F1评分：0.902 vs 0.789；所有$P < 0.001$）、基于AIC的LR（准确度：0.945 vs 0.937；AUC：0.904 vs 0.885；F1评分：0.864 vs 0.841；所有$P < 0.001$）、八特征SVM-RFE（准确度：0.961 vs 0.958，$P = 0.025$；AUC：0.940 vs 0.935，$P = 0.045$；F1评分：0.909 vs 0.903，$P = 0.047$）和XGBoost（准确度：0.934 vs 0.908，$P = 0.004$；AUC：0.953 vs 0.952，$P = 0.630$；F1评分：0.896 vs 0.874，$P = 0.108$）的总体疗效[图6（a）～（d）]。然而，SMOTE的应用并不影响CatBoost对基因型错配的预测能力（准确度：0.961 vs 0.960；AUC：0.909 vs 0.908；F1评分：0.940 vs 0.939；所有$P > 0.05$）[图6（e）]。

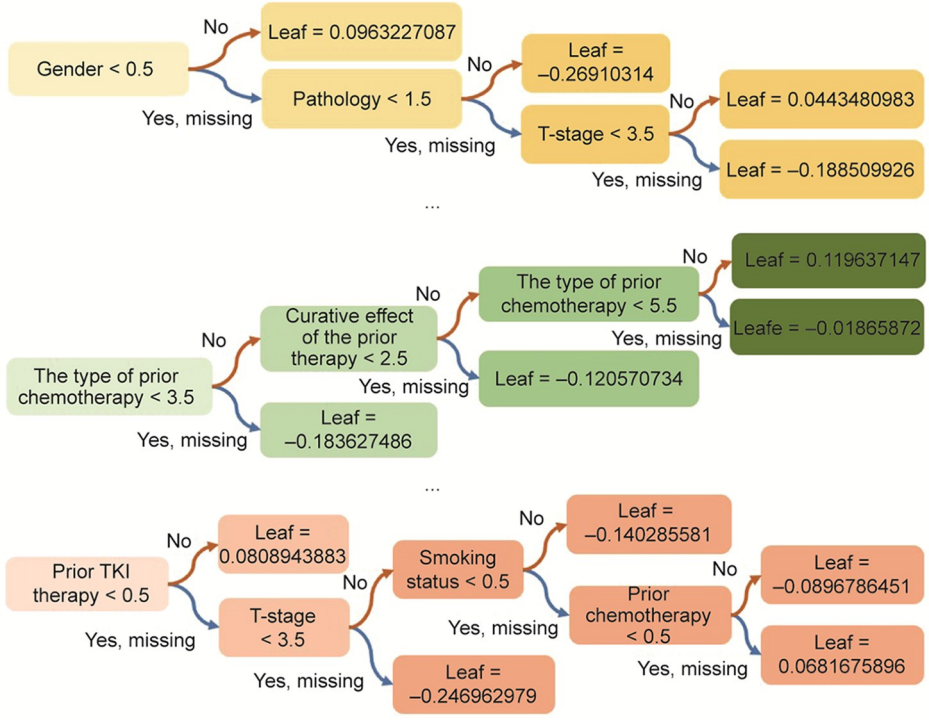
在这种情况下，CatBoost对均匀和不均匀样品均表现出相同的稳定潜力。然而，LR通过SMOTE实现了性能的显著增强，表明LR应该被推荐用于偶数数据。此外，本研究描述了一种可以在小的、不均匀的样本中改进SVM-RFE和XGBoost的方法。

Rank of importance	SVM-RFE	CatBoost	XGBoost
1	*EGFR* mutations	The number of driver gene mutations	The number of driver gene mutations
2	The number of driver gene mutations	*EGFR* mutations	*EGFR* mutations
3	Prior TKI therapy	The number of metastatic sites	Gender
4	Source	The type of prior chemotherapy	Curative effect of prior therapy
5	Gender	Age	Age
6	M-stage	Pathology	T-stage
7	T-stage	T-stage	Smoking status
8	Smoking status	Curative effect of the prior therapy	Prior TKI therapy
9	TNM-stage	N-stage	Pathology
10	Prior therapy	TNM-stage	Prior chemotherapy

(a)

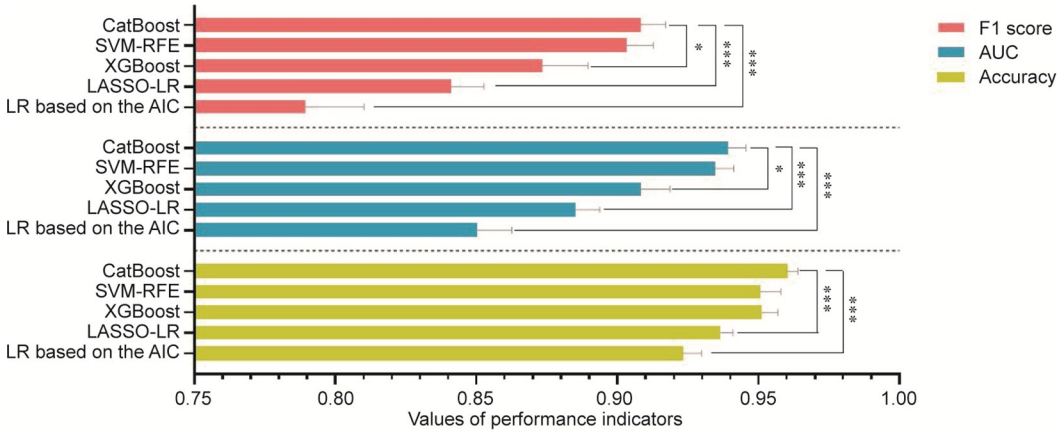


(b) (c)



(d)

图3. SVM-RFE和GBDT的前十位变量和建模过程的重要性等级。（a）根据三种算法（SVM-RFE、CatBoost和XGBoost）显示10个最关键变量的图表，其中相同的颜色代表相同的等级。（b）基于100个试验组中不同数量的变量，SVM-RFE的平均预测准确度。八特征SVM-RFE的准确度最高，变量最少。（c）SVM-RFE的平均F1评分是基于100个试验组中不同数量的变量。八特征SVM-RFE的F1评分最高，变量最少。（d）通过XGBoost模型训练获得的100个分类和回归树（CART）中的三个。将测试样本输入每个CART后，可以在叶节点获得每个样本的预测分数。权衡100个树的总分后，可以得到每个样本的总分和相应的分类。



(a)

Model	Index	Training group	Testing group	Bias between training groups and testing groups
LASSO-LR	Accuracy	0.947	0.937	0.010
	F1 score	0.861	0.841	0.020
	AUC	0.904	0.885	0.019
LR based on the AIC	Accuracy	0.966	0.923	0.043
	F1 score	0.905	0.789	0.116
	AUC	0.922	0.850	0.072
Eight-feature SVM-RFE	Accuracy	0.975	0.958	0.017
	F1 score	0.935	0.903	0.032
	AUC	0.973	0.951	0.022
XGBoost	Accuracy	0.981	0.951	0.030
	F1 score	0.950	0.874	0.076
	AUC	0.955	0.908	0.047
CatBoost	Accuracy	0.997	0.970	0.027
	F1 score	0.974	0.918	0.056
	AUC	0.977	0.949	0.028

(b)

图4. 不同模型之间的性能比较。（a）单一模型的性能。根据预测精度、AUC和100个测试组的F1评分， CatBoost和八特征SVM-RFE比基于最低AIC的XGBoost、LASSO-LR和LR表现出更好的性能。（b）训练组和测试组模型性能的偏倚。*$P < 0.05$、**$P < 0.01$、***$P < 0.001$，基于CatBoost和其他模型之间的配对样本t检验。

Rank of importance	SMOTE + SVM-RFE	SMOTE + CatBoost	SMOTE + XGBoost
1	The number of driver gene mutations	The number of driver gene mutations	The number of driver gene mutations
2	*EGFR* mutations	*EGFR* mutations	T-stage
3	Prior TKI therapy	Smoking status	*EGFR* mutations
4	The number of metastatic sites	Prior chemotherapy	Gender
5	N-stage	Age	Smoking status
6	The type of prior chemotherapy	Pathology	The number of metastatic sites
7	Curative effect prior therapy	T-stage	Prior chemotherapy
8	Source	Curative effect of the prior therapy	The type of prior chemotherapy
9	Smoking status	N-stage	Prior TKI therapy
10	Gender	TNM-stage	Curative effect of prior therapy

图5. 根据三种算法（SVM-RFE、CatBoost和XGBoost）使用SMOTE的10个最关键变量，其中相同的颜色代表相同的排序。

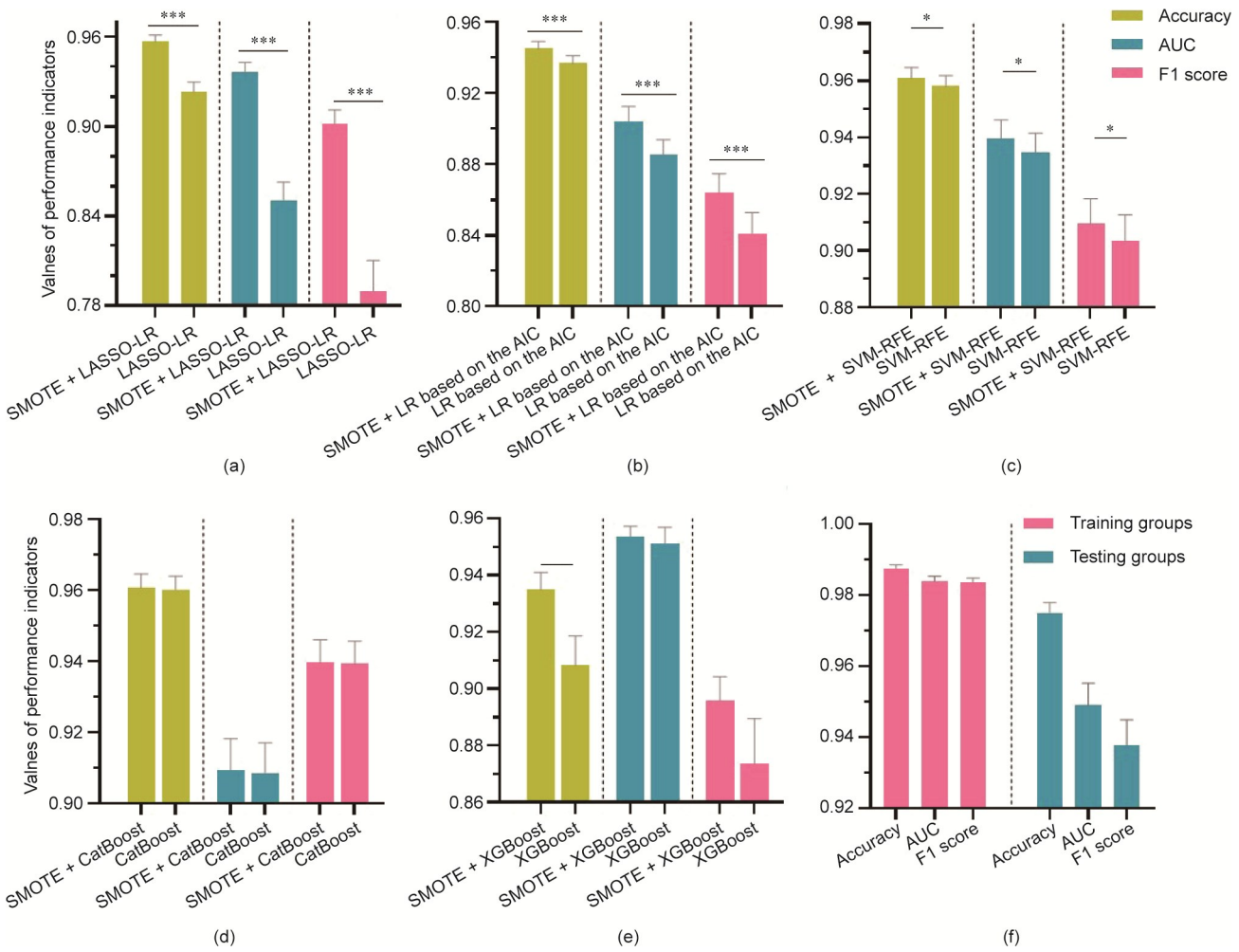


图6. 不同算法的SMOTE的性能。（a）在LASSO-LR中采用SMOTE的性能。（b）根据AIC向LR引入SMOTE的性能。（c）将SMOTE引入八特征SVM-RFE的性能。（d）将SMOTE引入XGBoost的性能。（e）将SMOTE引入CatBoost的性能。（f）在通过进行基于AIC、八特征SVM-RFE、XGBoost和CatBoost的LR的SMOTE的训练组与测试组中集成分类器的性能。*$P < 0.05$、**$P < 0.01$、***$P < 0.001$，基于配对样本t检验。

3.5.3. 性能集成分类器的最终优化

考虑到应用SMOTE后大多数模型性能的急剧增强，本研究最终在基于AIC、八特征SVM-RFE、XGBoost和CatBoost的LR的基础上进行SMOTE后使用了集成分类器。令人惊讶的是，与单一模型相比，集成分类器的准确度（平均值= 0.975）、AUC（平均值= 0.949）和F1评分（平均值= 0.938）得到了进一步的优化[图6（f）]。而且，训练组和测试组集成分类器的偏倚也更优（所有差异均在5%以下）[图6（f）]。

因此，本文提出了一种基于单一优化模型的集成分类器，克服了样本量和分布缺陷，达到了最佳的辨别和稳定性水平。

4. 讨论和结论

本研究最初开发了NOG/PDX模型和患者样本之间驱动基因突变不一致的预测模型。共有53个肺癌NOG/PDX模型被成功植入，包括42个驱动基因突变与亲代肿瘤相匹配的NOG/PDX模型和11个不匹配肿瘤的NOG/PDX模型。为了分析这个不平衡数据库，本研究使用了5种算法：基于AIC的LR、LASSO-LR、SVM-RFE、XGBoost和CatBoost。根据测试组中的指标，CatBoost和SVM-RFE的性能最好。此外，使用SMOTE改善了除CatBoost以外的所有模型在基本水平上的性能。最后，基于单一模型的集成分类器性能最好（平均准确度= 0.975；平均AUC = 0.949；平均F1分数= 0.938），训练组和测试组之间的偏倚可以接受（所有差异均在5%以下）。

PDX模型的生成和传代是经常发生克隆和亚克隆改变的动态事件，尤其是当P1 PDX模型的发展缓慢时，这为肿瘤细胞提供了足够的时间进行突变并适应新的环境[27–28]。除了细胞自主异质性，肿瘤微环境（TME）中的基质异质性是PDX驱动基因型不同于亲代肿瘤的一个

关键原因[12]。据报道，SCC在裸鼠中比ADC更容易发生肿瘤 [18]，这与本研究的结论不一致，即SCC是用基因匹配建立NOG/PDX模型最具挑战性的肿瘤类型。在SCC肿瘤中检测到的$CD8^+$ TIL多于在非SCC细胞巢中检测到的$CD8^+$ TIL [29]，表明SCC的PDX模型在异种移植物植入过程中可能会丢失更多的肿瘤基质。此外，已发现SCC携带的克隆突变明显多于ADC [30]，这有助于进行更多的克隆选择。尽管年龄在多变量LR中的权重较小，但本研究尚未发现适当的方法用来说明年龄较小而非年龄较大是驱动基因匹配的风险因素[31]。大多数PDX模型使用8周龄小鼠而非老年小鼠（> 8个月），而最近的研究发现，衰老可显著改变TME的组分[32]。因此，小鼠和患者年龄的不一致可能是发现年龄是此处预测特征的原因。另一个特征（来源）在基因型匹配中也发挥了负面作用，这与肿瘤植入不同。尽管液体来源被认为具有较高的植入率[33]，但本研究发现在液体来源的肿瘤异种移植物中维持亲代肿瘤的驱动基因型更具挑战性。

驱动基因突变的数量（包括克隆和亚克隆突变）与肿瘤内异质性、基因组不稳定性和染色体不稳定性相关[34]。首先，多变量LR模型中驱动基因数量的最大系数也说明了其在开发非患者匹配基因型中的重要性。其次，据报道，*EGFR*突变肺癌的PDX模型的组织分化差，*EGFR*突变丢失频繁[35]；这支持了本研究中NOG/PDX模型中*EGFR*突变不一致的高风险。再次，有证据表明培美曲塞可增加TIL的数量，并上调与抗原呈递相关的免疫相关基因；这可能支持以下结论：来自接受培美曲塞治疗患者的PDX模型不太可能维持原始基因型[36]。已证实TKI可改变肺部的TME，包括$CD8^+$ T细胞和单核髓源性抑制细胞（M-MDSC; $CD11b^+Ly6^-G^-Ly6C^{high}$）增加，以及$Foxp3^+$ T调节细胞（Tregs）和M2样巨噬细胞（$CD206^+$）减少[37]。此外，在TKI治疗期间经常发生克隆选择，导致TKI耐药[38]。有趣的是，研究发现在NOG/PDX模型建立过程中促进TIL的因子促进了基因型的稳定性，但这需要进一步的验证[图7（a）]。

最近，ML已成为许多领域预测建模的有用方法，因为该方法使预测模型能够从初始数据中系统地“学习”信息，并适应每个新的数据环境[39]。然而，ML尚未被广泛应用于小样本数据库（每个预测变量少于10个频率），这是生物医学动物模型的共同特征，即成本高、技术复杂[40]。最终，用来建立肺癌NOG/PDX模型预测工具的ML算法具有优异的性能，该算法不仅为用于筛选肺癌患者进行精确免疫治疗的NOG/PDX模型提供了预测工具，同时也为构建生物医学样本较小的预测模型提供了通用的方法[图7（b）]。

本研究仍存在一些局限性。首先，纳入研究的患者数量有限，应开展更大规模的实验来进一步验证这些结论。其次，由于训练数据有限，该模型的预测结果不能准确到每个驱动基因突变。第三，*EGFR*突变状态既是自变量，也是结果，可能引起共线性。最后，潜在的选择偏倚是不可避免的，这项研究的性别比是不均匀的。应进行更大规模的试验来进一步验证这些结论。

综上，本研究建立了基于ML的NOG/PDX模型与患者样本驱动基因突变不一致性的预测模型，有望提高PDX建立的成功率，减少巨大的经济损失。此外，尽管NOG小鼠没有得到很好的研究，但本研究中使用的NOG小鼠被认为是构建癌症免疫治疗PDX模型的极好选择。因此，本研究建立的模型具有免疫治疗筛选和开发的潜力。

致谢

图5是用BioRender.com创建的。语言润色由Springer Nature Author Services（SNAS）的本地编辑完成。本研究得到了国家自然科学基金项目（81802255）的部分资助；感谢上海市肺科医院临床研究项目（FKLY20010、FKLY20001、fk18005）、上海青年人才（2019 QNBJ）梦想导师“杰出优秀人才计划”（fkyq1901）、2019年重点学科（肿瘤学）、上海市卫生健康委员会项目（201940192）、上海市肺科医院科研项目（fkcx1903）、上海市卫生和计划生育委员会（2017YQ050）、同济大学SITP创新培养项目——领军人才重点项目（19411950300）、上海市医院协会医院管理研究基金青年项目（Q1902037）的资助，以及感谢上海市临床重点专科建设项目呼吸内科肺非传染性疾病多学科协作体系的推广应用。

Authors’ contributions

Yayi He, Xuzhen Tang, Junjie Zhu, and Yang Yang conceived the concepts of the research. Yayi He, Haoyue Guo, and Li Diao designed the research. Hui Qi and Chunlei Dai established PDX models. Haoyue Guo performed the image preprocessing. Haoyue Guo and Li Diao contributed to machine learning. All authors analyzed the data and wrote the paper.

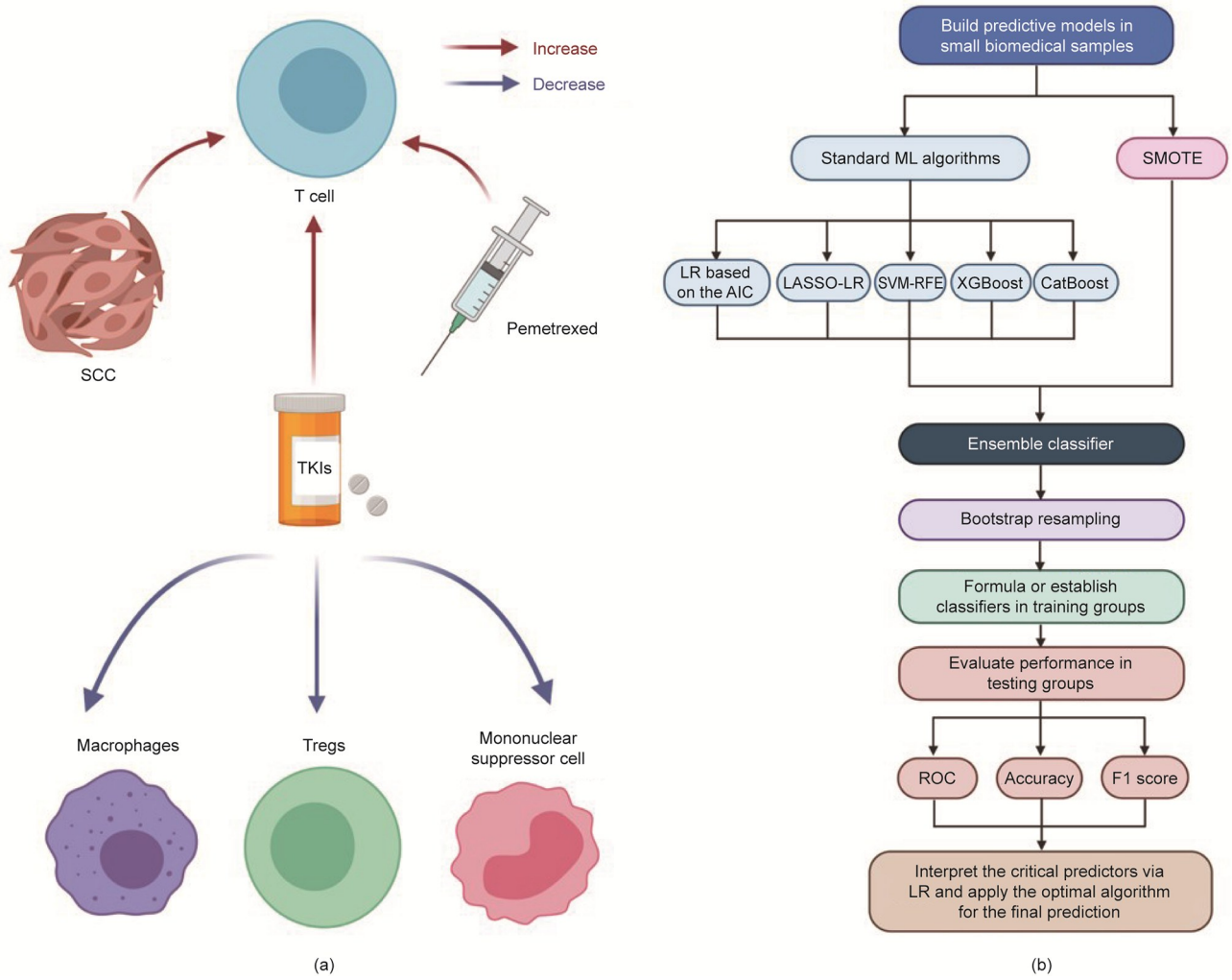


图7. TME对驱动基因突变的潜在影响和在小数据集中构建预测模型的流程图。（a）导致驱动基因突变不一致的因素与TME之间的相关性。根据单变量和多变量LR，SCC、培美曲塞应用和既往TKI治疗是PDX模型和亲代肿瘤之间不匹配基因型的风险因素。所有这三个因素均引起TIL。而且，TKI可以减少TME中$Foxp3^+$ Tregs、单核髓源性抑制细胞（$CD11b^+Ly6^-G^-Ly6C^{high}$）和M2样巨噬细胞（$CD206^+$）。（b）在小型生物医学数据集中建立预测模型的流程图：①当数据集不均匀时，首先进行SMOTE。使用标准ML算法选择特征，在所有样本中开发多变量模型，包括基于AIC、LASSO-LR、SVM-RFE、XGBoost、CatBoost等的逐步LR。②提出一种基于优化的单一模型的集成分类器。③进行自举重新采样，避免过度拟合，达到性能稳定。④制定预测评分或建立训练组中的预测分类器。⑤根据相应试验组的ROC、准确度和F1评分评价预测模型，确定最佳的建模算法。⑥通过LR解释正类的关键预测因子，并应用最优算法进行最终预测。

Compliance with ethics guidelines

Yayi He, Haoyue Guo Li Diao, Yu Chen, Junjie Zhu, Hiran C. Fernando, Diego Gonzalez Rivas, Hui Qi, Chunlei Dai, Xuzhen Tang, Jun Zhu, Jiawei Dai, Kan He, Dan Chan, and Yang Yang declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2021.06.017.

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021;71(1):7–33.
[2] Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. Transl Lung Cancer Res 2016;5(3):288–300.
[3] Politi K, Herbst RS. Lung cancer in the era of precision medicine. Clin Cancer Res 2015;21(10):2213–20.
[4] Reck M, Rodríguez–Abreu D, Robinson AG, Hui R, Cs}oszi T, Fülöp A, 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(7):537–46.
[5] Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, Tsurutani J, 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(2):121–8.
[6] Rosell R, Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, 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(3):239–46.

[7] Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448(7153):561–6.

[8] Zhao ZR, Wang JF, Lin YB, Wang F, Fu S, Zhang SL, et al. Mutation abundance affects the efficacy of *EGFR* tyrosine kinase inhibitor readministration in nonsmall-cell lung cancer with acquired resistance. Med Oncol 2014;31(1):810.

[9] Lim ZF, Ma PC. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. J Hematol Oncol 2019;12(1):134.

[10] Liang W, Guo M, Pan Z, Cai X, Li C, Zhao Y, 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(6):2014–21.

[11] Cutz JC, Guan J, Bayani J, Yoshimoto M, Xue H, Sutcliffe M, 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(13):4043–54.

[12] Cassidy JW, Caldas C, Bruna A. Maintaining tumor heterogeneity in patientderived tumor xenografts. Cancer Res 2015;75(15):2963–8.

[13] Chijiwa T, Kawai K, Noguchi A, Sato H, Hayashi A, Cho H, et al. Establishment of patient-derived cancer xenografts in immunodeficient NOG mice. Int J Oncol 2015;47(1):61–70.

[14] Ny L, Rizzo LY, Belgrano V, Karlsson J, Jespersen H, Carstam L, et al. Supporting clinical decision making in advanced melanoma by preclinical testing in personalized immune-humanized xenograft mouse models. Ann Oncol 2020;31(2):266–73.

[15] Jespersen H, Lindberg MF, Donia M, Söderberg EMV, Andersen R, Keller U, et al. Clinical responses to adoptive T-cell transfer can be modeled in an autologous immune-humanized mouse model. Nat Commun 2017;8(1):707.

[16] Fichtner I, Rolff J, Soong R, Hoffmann J, Hammer S, Sommer A, et al. Establishment of patient-derived non-small cell lung cancer xenografts as models for the identification of predictive biomarkers. Clin Cancer Res 2008; 14(20):6456–68.

[17] John T, Kohler D, Pintilie M, Yanagawa N, Pham NA, Li M, et al. 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(1):134–41.

[18] Zhang XC, Zhang J, Li M, Huang XS, Yang XN, Zhong WZ, et al. Establishment ofpatient-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(1):168.

[19] Izumchenko E, Paz K, Ciznadija D, Sloma I, Katz A, Vasquez-Dunddel D, et al. Patient-derived xenografts effectively capture responses to oncology therapy in a heterogeneous cohort of patients with solid tumors. Ann Oncol 2017;28(10): 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 Inform 2018;16(2):30–5.

[21] Hidalgo M, Amant F, Biankin AV, Budinská E, Byrne AT, Caldas C, et al. Patientderived xenograft models: an emerging platform for translational cancer research. Cancer Discov 2014;4(9):998–1013.

[22] Park B, Jeong BC, Choi Y-L, Kwon GY, Lim JE, Seo SI, et al. Development and characterization of a bladder cancer xenograft model using patient-derived tumor tissue. Cancer Sci 2013;104(5):631–8.

[23] Xu H, Zhao X, Shi Y, Li X, Qian Y, Zou J, et al. Development and validation of a simple-to-use clinical nomogram for predicting obstructive sleep apnea. BMC Pulm Med 2019;19(1):1.

[24] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321–57.

[25] Tomaschek F, Hendrix P, Baayen RH. Strategies for addressing collinearity in multivariate linguistic data. J Phonetics 2018;71:249–67.

[26] Sain H, Purnami SW. Combine sampling support vector machine for imbalanced data classification. Procedia Comput Sci 2015;72:59–66.

[27] McFadden D, Papagiannakopoulos T, Taylor-Weiner A, Stewart C, Carter S, Cibulskis K, et al. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. Cell 2014;156(6):1298–311.

[28] Fu S, Zhao J, Bai H, Duan J, Wang Z, An T, et al. High-fidelity of non-small cell lung cancer xenograft models derived from bronchoscopy-guided biopsies. Thorac Cancer 2016;7(1):100–10.

[29] Meng X, Gao Y, Yang L, Jing H, Teng F, Huang Z, et al. Immune microenvironment differences between squamous and non-squamous nonsmall-cell lung cancer and their influence on the prognosis. Clin Lung Cancer 2019; 20(1):48–58.

[30] Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. TRACERx Consortium. Tracking the evolution of non-small-cell lung cancer. N Engl J Med 2017;376(22):2109–21.

[31] Milholland B, Auton A, Suh Y, Vijg J. Age-related somatic mutations in the cancer genome. Oncotarget 2015;6(28):24627–35.

[32] Fane M, Weeraratna AT. How the ageing microenvironment influences tumour progression. Nat Rev Cancer 2020;20(2):89–106.

[33] Mattar M, McCarthy CR, Kulick AR, Qeriqi B, Guzman S, de Stanchina E. Establishing and maintaining an extensive library of patient-derived xenograft models. Front Oncol 2018;8:19.

[34] Raynaud F, Mina M, Tavernari D, Ciriello G, Kool M. Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. PLoS Genet 2018;14(9):e1007669.

[35] 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(1):139.

[36] Novosiadly R, Schaer D, Lu Z, Amaladas N, Luo S, Capen A, et al. P3.07-006 pemetrexed exerts intratumor immunomodulatory effects and enhances efficacy of immune checkpoint blockade in MC38 syngeneic mouse tumor model. J Thorac Oncol 2017;12(11):S2300.

[37] Jia Y, Li X, Jiang T, Zhao S, Zhao C, Zhang L, et al. *EGFR*-targeted therapy alters the tumor microenvironment in *EGFR*-driven lung tumors: implications for combination therapies. Int J Cancer 2019;145(5):1432–44.

[38] Wang F, Diao XY, Zhang X, Shao Q, Feng YF, An X, et al. Identification of genetic alterations associated with primary resistance to *EGFR*-TKIs in advanced nonsmall-cell lung cancer patients with *EGFR* sensitive mutations. Cancer Commun 2019;39(1):7.

[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(20):469–74.

[40] Cohen ME, Hudson DL. New chaotic methods for biomedical signal analysis. In: Naguib RNG, Solaiman B, Nagy G, Le Guillou C, Roa L, Beltrame F, editors. Proceedings of the Proceedings 2000 IEEE EMBS International Conference onInformation Technology Applications in Biomedicine ITAB-ITIS 2000 Joint Meeting Third IEEE EMBS International Conference on Information Technol; 2000 Nov 9–10; Arlington, TX, USA. York City: Curran Associates; 2002.