

基于机器学习算法的自体外周血造血干细胞采集预测模型构建与应用

李若冰 唐古生 罗艳蓉 黄佳莹 张倩倩 鲁桂华[△]

(海军军医大学第一附属医院血液科 上海 200433)

【摘要】 目的 筛选自体外周血造血干细胞(peripheral blood stem cell, PBSC)采集的危险因素并建立个体风险预测模型,以提高临床中自体PBSC采集的成功率。**方法** 通过大数据平台收集2013年2月至2021年5月在海军军医大学第一附属医院血液科行自体PBSC采集术的恶性血液病患者757例,对患者进行单因素显著性统计学分析和多因素Logistic回归分析对PBSC采集危险因素进行筛选。采用Python 3.8.8版本、PyCharm 2021.1.3集成开发环境构建Logistic回归模型和前馈神经网络、最小二乘支持向量机、自动机器学习3种机器学习模型,并采用多种模型评价指标对其评价。**结果** 共收集患者PBSC采集前的指标24项,单因素和Logistic回归分析筛选出11项PBSC采集危险因素。所构建的Logistic回归模型、前馈神经网络、最小二乘支持向量机和自动机器学习模型对自体PBSC采集风险预测的准确度分别为0.822、0.873、0.875和0.973。**结论** 本研究自建自动机器学习模型能够准确预测自体PBSC采集结果,对提高临床自体PBSC采集成功率具有重要参考价值。

【关键词】 血液病; 造血干细胞(PBSC); 机器学习; 预测模型

【中图分类号】 TP399, R552 **【文献标志码】** A **doi:** 10.3969/j.issn.1672-8467.2023.03.011

Prediction model construction and application of machine learning algorithms for outcome prediction in autologous peripheral blood hematopoietic stem cell collection

LI Ruo-bing, TANG Gu-sheng, LUO Yan-rong, HUANG Jia-ying, ZHANG Qian-qian, LU Gui-hua[△]

(Department of Hematology, First Affiliated Hospital of Naval Medical University, Shanghai 200433, China)

【Abstract】 Objective To screen risk factors for autologous peripheral blood hematopoietic stem cell (PBSC) collection and develop an individual risk prediction model to improve the success rate of autologous PBSC collection in the clinic. **Methods** A total of 757 patients with hematologic malignancies who underwent PBSC collection in the Department of Hematology, First Affiliated Hospital of Naval University from Feb 2013 to May 2021 were collected through the big data platform, and the patients were screened for risk factors of PBSC collection by univariate statistical analysis and multivariate Logistic regression. Logistic regression models and three machine learning models, BP neural network (BPNN), least squares support vector machine (LSSVM), and automated machine learning (Auto-ML), were constructed using Python version 3.8.8, Pycharm 2021.1.3 integrated development environment, and the models were evaluated using several model evaluation metrics. **Results** A total of 24 items of the index before PBSC collection from patients were collected, and 11 items with risk factors

上海市自然科学基金(20ZR1457000)

[△]Corresponding author E-mail: Lovelugh@163.com

网络首发时间:2023-03-22 11:18:46 网络首发地址:https://kns.cnki.net/kcms/detail/31.1885.R.20230321.0928.002.html

for PBSC collection were screened by univariate and logistic regression analysis. The accuracies of the constructed logistic, BPNN, LSSVM and Auto-ML models for risk prediction of autologous PBSC collection were 0.822, 0.873, 0.875 and 0.973, respectively. **Conclusion** The established Auto-ML model can accurately predict the outcome of autologous PBSC collection and will be valuable for improving the success rate of autologous PBSC collection in the clinic.

【Key words】 hematopathy; hematopoietic stem cell (PBSC); machine learning; prediction model

* This work was supported by the Natural Science Foundation of Shanghai (20ZR1457000).

自体外周血造血干细胞移植 (autologous peripheral blood hematopoietic stem cell transplantation, ASCT) 是淋巴瘤、多发性骨髓瘤等多种恶性血液病的重要治疗手段^[1-3], 尤其是在我国少子女或单子女家庭逐渐普及的情况下, 人类白细胞抗原相合同胞做供者的机会越来越少。ASCT 一般是通过离心法先将患者自体外周血造血干细胞 (peripheral blood stem cell, PBSC) 进行提取并冷冻保存, 然后对患者进行大剂量化疗和 (或) 放疗后再回输给患者, 使之快速重建免疫系统^[4-5]。人类 PBSC 在外周血中含量较少, 约占单个核细胞的 0.1%~1.0%, 通常需要采用离心法对自体 PBSC 进行提纯, 以获得一定数量的高纯度自体 PBSC^[6]。但是, 行 PBSC 术时如果无法获取足够的 PBSC 则会影响后期回输免疫系统的重建效果, 并最终造成 ASCT 失败, 严重威胁血液病患者生存, 因此 PBSC 采集成功是 ASCT 的先决条件^[7-8]。受患者个体差异的影响, 临床上 PBSC 采集失败案例时有发生, 造成患者需要多次采集才能达到 ASCT 移植标准, 这给患者及家属造成严重的心理压力和沉重的经济负担^[9]。如果患者在行自体 PBSC 采集术之前能够对 PBSC 采集风险进行预测, 则医护人员便能够根据预测风险信息对 PBSC 采集术进行修正, 以此提高自体 PBSC 采集的成功率^[10]。目前对自体 PBSC 采集风险预测的相关研究鲜有报道, 医护人员多依靠自己的工作经验对患者进行评估, 以此来制定相应的 PBSC 采集方案^[11-12]。然而, 受医护人员的教育水平和从业经验个体差异的影响, 往往无法准确预测 PBSC 结果^[9, 13]。因此探索一种具有普适性的自体 PBSC 采集风险的预测方法迫在眉睫。将人工智能技术应用于 PBSC 的采集, 可能有效提高临床对自体 PBSC 采集结果的预测精度。本研究从数据驱动的角度出发, 使用机器学习与统计学相关理论方法, 首次成功利用前馈神经网络 (back

propagation neural network, BPNN)、最小二乘支持向量机 (least square support vector machine, LSSVM) 和自动机器学习 (automated machine learning, Auto-ML) 等人工智能技术, 构建了血液病患者 PBSC 采集结果风险预测模型, 分别采用单因素和多因素相关性分析手段对 24 项原始变量进行筛选, 利用筛选后的特征变量数据对所建模型进行训练和验证, 并对模型输入的相关特征变量进行讨论分析, 以期临床医护人员的采集工作提供指导, 并提高 PBSC 采集成功率。

资 料 和 方 法

研究对象 选取 2013 年 02 月至 2021 年 05 月在海军军医大学第一附属医院血液科行自体 PBSC 采集术的恶性血液病患者作为研究对象, 恶性血液病类型包括: 急性淋巴细胞白血病 (acute lymphoblastic leukemia, ALL)、多发性骨髓瘤 (multiple myeloma, MM)、非霍奇金淋巴瘤 (non-Hodgkin lymphoma, NHL)、霍奇金淋巴瘤 (Hodgkin's lymphoma, HL)。纳入标准: (1) 年龄 ≥ 12 周岁。(2) 首次进行自体 PBSC 采集。(3) 动员方案相同, 采集时间间隔 4~6 天。排除标准: (1) 临床资料不全者; (2) 存在交流障碍患者。根据纳入标准初步选择 793 例患者为研究对象, 按照排除标准剔除 36 例资料不全或存在交流障碍患者, 最终纳入 757 例血液病患者作为研究对象, 筛选前后患者基线资料无差异, 且在本研究执行过程中已对所有数据进行了脱敏处理。本研究通过我院医学伦理委员会审批 (批准号: CHEC2022-076)。

PBSC 采集和计数 所有患者均采用重组人粒细胞刺激因子 (granulocyte colony stimulating factor, G-CSF), 动员剂量 $5\sim 10\ \mu\text{g}\cdot\text{kg}^{-1}\cdot\text{d}^{-1}$, 连续注射 4~6 天, 每天皮下注射 1 次, 注射时间为上午 9:00—

9:30。在动员第4天开始评估患者白细胞和单个核细胞(包括单核细胞和淋巴细胞)比例,以确定采集的最佳时机。采用德国 Fresenius 血细胞分离机 COM-ETC 的单个核细胞采集程序进行采集,采集前用 0.9% 生理盐水预冲管道,用复方枸橼酸钠溶液抗凝,并予地塞米松 10 mg 静脉推注。循环总量设为 3.5~4 倍的全身血容量,流速为 50~60 mL/min,全血与抗凝剂流速比为 10:1~12:1,其余参数为系统默认值。考虑到多数患者在采集开始后会出现不同程度的低钙,在 PBSC 采集开始后,给予患者 10% 葡萄糖酸钙 60 mL,经回输管道以 10~15 mL/h 注射泵泵入。

在循环结束后分别对 CD34⁺ 和单个核细胞(mono-nuclear cell, MNC)进行计数。每次 CD34⁺ 检测均设置对照组,取 PBSC 采集物 1 mL 并调整细胞数至 $(0.5\sim1.0)\times10^6/\text{mL}$,取 50 μL ,加入适量各种荧光抗体标记,温室避光孵育 15 min,加入裂红液 2 mL,4 000 $\times g$ 离心 5 min,弃上清,加入 PBS 缓冲液 2 mL,4 000 $\times g$ 离心 5 min,每管加入 PBS 缓冲液约 0.3 mL,以 200 目尼龙膜过滤后上机检测。检测设备采用美国 FACSCanto II 型流式细胞仪和 FACSDiva 软件,使用 CD34-PE、CD45-PerCP、FSC 和 SSC 等 4 个参数,累积设门,分母为 CD45⁺ WBC,检测至 2×10^5 个有核细胞,取实验组和检测组 CD34⁺ 细胞群含量平均值作为患者最终 CD34⁺ 细胞数。

同时制作 PBSC 采集物涂片两张,采用瑞氏吉姆萨染液(珠海贝索生物技术有限公司)染色,在显微镜下对 MNC 细胞进行分类计数,选择两张图片 MNC 计数均值作为最终患者 MNC 细胞计数。

特征选取 根据淋巴瘤诊疗指南^[14]、中国多发性骨髓瘤诊治指南(2020 年修订)^[15]及自体 PBSC 采集相关研究^[16],收集患者临床基本资料,包括性别、年龄、BMI、吸烟、血液、钾、钙、钠、患病类型等 9 项一般指标和患者行 PBSC 采集术前的 C 反应蛋白、白细胞计数、淋巴细胞、单核细胞、红细胞、血小板、血红蛋白等 15 项血常规指标。

统计学方法 采用 SPSS 25.0 和 R4.1.2 进行统计学分析。符合正态分布的计量资料以 $\bar{x}\pm s$ 表示,两组组间差异采用独立样本 t 检验;如不符合正态分布的计量资料用 $M(P_{25}, P_{75})$ 表示,两组组间采用非参数秩和检验。计数资料以频数和百分比表示,组间比较采用 χ^2 检验。PBSC 采集成功相关危险因

素分析采用多因素 Logistic 回归,结果以 OR 和 95% CI 表示并进行描述性分析,检验水准 $\alpha=0.05$ 。

机器学习模型构建 根据 2018 版造血干细胞移植治疗淋巴瘤中国专家共识^[17]及本院专家经验, PBSC 采集成功标准定为 CD34⁺ 细胞计数 $\geq 2\times10^6/\text{kg}$ 且 MNC 细胞计数 $\geq 5\times10^8/\text{kg}$ 。将 757 例患者分为采集成功组和采集失败组,对两组患者的 24 项指标进行显著性统计学分析及多因素 Logistic 回归分析。以 757 例患者经 Logistic 回归分析有意义的所有指标纳入机器学习模型,采用 5 折交叉验证法对机器学习模型进行训练和验证。本研究采用 Logistic 回归, BPNN 模型、LSSVM 模型和 Auto-ML 模型分别对 PBSC 采集结果进行预测。上述模型均采用 Python 3.8.8 版本、PyCharm 2021.1.3 集成开发环境来构建,最后通过敏感度、特异度、准确度和 AUC 对 4 种模型进行评价。

结 果

患者一般资料 757 例行自体 PBSC 采集术患者的一般资料见表 1,其中采集成功组 592 例,采集失败组 165 例,两组患者的性别和吸烟史差异有统计学意义。患者的血常规资料见表 2,两组之间 C 反应蛋白、白细胞计数、红细胞计数、平均红细胞血红蛋白含量、平均红细胞血红蛋白浓度、血红蛋白含量、红细胞比积、血小板计数、血小板分布宽度差异均有统计学意义。

Logistic 回归分析 根据上述统计学分析结果,将 11 项单因素分析中 $P<0.05$ 的因素进行逐步向前多因素 Logistic 回归分析(表 3),以自体 PBSC 是否采集成功为因变量(是=1,否=0)。结果显示:患者性别、吸烟史、C 反应蛋白、白细胞计数、红细胞计数、平均红细胞血红蛋白含量、平均红细胞血红蛋白浓度、血红蛋白含量、红细胞比积、血小板计数、血小板分布宽度与 PBSC 采集存在相关性。

机器学习模型 将表 1 和 2 中采集成功组和采集失败组比较有统计学差异的 11 项指标纳入 4 种机器学习模型,并通过 5 折交叉验证法对训练集进行训练。表 4 为 Logistic、BPNN、LSSVM、Auto-ML 模型的 4 种性能评价指标,可以看出 BPNN、LSSVM、Auto-ML 模型的灵敏度、特异度、准确度和 AUC 均高于 Logistic 模型,可以认为机器学习算

表1 采集成功组与采集失败组一般资料比较

Tab 1 Comparison of general data between collection success group and collection failure group [$\bar{x} \pm s$ or $n(\%)$]

Characteristic	PBSC success group ($n=592$)	PBSC failure group ($n=165$)	t/χ^2	P
Age (y)	47.13 \pm 12.37	46.05 \pm 12.18	1.003 ⁽¹⁾	0.158
Gender			17.542 ⁽²⁾	<0.001
Male	310 (52.37)	56 (33.94)		
Female	282 (47.63)	109 (66.06)		
BMI (kg/m ²)	23.29 \pm 3.53	23.04 \pm 3.78	0.762 ⁽¹⁾	0.223
Smoke (y)	13.36 \pm 4.05	15.65 \pm 6.02	-4.625 ⁽¹⁾	<0.001
SBP (mmHg)	137.24 \pm 13.45	136.51 \pm 14.82	0.571 ⁽¹⁾	0.284
DBP (mmHg)	86.25 \pm 9.59	85.37 \pm 8.18	1.175 ⁽¹⁾	0.120
K (mmol/L)	3.83 \pm 0.35	3.78 \pm 0.37	1.553 ⁽¹⁾	0.060
Ca (mmol/L)	2.25 \pm 0.13	2.27 \pm 0.14	-1.648 ⁽¹⁾	0.128
Na (mmol/L)	142.88 \pm 4.13	142.65 \pm 2.33	0.725 ⁽¹⁾	0.235
Disease type			1.476 ⁽²⁾	0.688
MM	302 (51.01)	84 (50.91)		
HL	48 (8.11)	18 (10.91)		
NHL	196 (33.11)	50 (30.30)		
ALL	46 (7.77)	13 (7.88)		

BMI: Body mass index; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; ALL: Acute lymphoblastic leukemia; MM: Multiple myeloma; NHL: Non-Hodgkin lymphoma; HL: Hodgkin's lymphoma. ⁽¹⁾ t test; ⁽²⁾ χ^2 test.

表2 采集成功组与采集失败组血常规指标比较

Tab 2 Comparison of blood routine indexes between collection success group and collection failure group [$\bar{x} \pm s$]

Characteristic	PBSC success group ($n=592$)	PBSC failure group ($n=165$)	t	P
CRP (mg/L)	3.75 \pm 1.20	3.11 \pm 1.33	5.580	<0.001
Leukocyte ($\times 10^9/L$)	16.27 \pm 10.69	12.08 \pm 7.14	5.914	<0.001
Lymphocyte (%)	12.39 \pm 7.66	12.31 \pm 9.52	0.099	0.461
Monocyte (%)	19.30 \pm 14.18	19.25 \pm 8.64	0.056	0.448
Neutrophil ($\times 10^9/L$)	4.37 \pm 1.88	4.34 \pm 1.89	0.181	0.428
Eosinophil ($\times 10^9/L$)	0.13 \pm 0.14	0.14 \pm 0.13	-0.859	0.196
Basophil ($\times 10^9/L$)	0.03 \pm 0.01	0.03 \pm 0.02	1.242	0.108
Erythrocyte ($\times 10^{12}/L$)	4.47 \pm 0.48	3.48 \pm 0.61	19.192	<0.001
MCH (pg)	31.76 \pm 2.53	31.32 \pm 1.8	2.522	0.006
MCHC (g/L)	336.42 \pm 12.21	333.05 \pm 11.24	3.340	<0.001
Hemoglobin (g/L)	111.32 \pm 15.64	100.11 \pm 16.82	7.685	<0.001
Hematocrit (%)	0.49 \pm 0.03	0.43 \pm 0.06	12.420	<0.001
Platelet ($\times 10^9/L$)	187.38 \pm 31.25	180.56 \pm 41.12	1.977	0.024
PDW (%)	13.41 \pm 1.46	12.57 \pm 2.01	5.012	<0.001
MPV (fl)	11.16 \pm 0.81	11.12 \pm 0.73	0.607	0.272

CRP: C-reactive protein; MCH: Mean corpuscular hemoglobin; MCHC: Mean corpuscular hemoglobin concentration; PDW: Platelet distribution width; MPV: Mean platelet volume.

法对自体PBSC采集结果有着较好的预测能力,且3种机器学习模型中Auto-ML最优。此外,通过对BPNN、LSSVM、Auto-ML模型11项输入参数的重要性进行分析,发现11项指标在3种机器学习模型中的重要性存在较大差异,主要反映在BPNN、LSSVM模型输入参数的重要性分布相对均匀,而Auto-ML模型输入参数的相对重要性主要集中在红细胞和血小板,经过计算发现红细胞和血小板的相对

重要性在Auto-ML模型中占比超过72.48%(图1)。

临床应用 为了进一步验证Auto-ML模型,以2021年7月至2022年3月在我院血液科行PBSC采集术的恶性血液病患者107例为研究对象,其中PBSC采集成功86例,PBSC采集失败21例;男性67例,女性40例;MM、HL、NHL和ALL患者分别为53、14、30、10例,与模型训练数据无明显差异。收集107例患者的以上11项指标作为模型输入参

表3 PBSC 采集差异性指标 Logistic 回归分析结果

Tab 3 Logistic regression analysis results of difference index in PBSC collection

Characteristic	β	SE	Ward χ^2	OR (95%CI)	P
Gender (%)	0.139	0.423	0.121	1.051 (0.259–6.101)	0.011
Smoke (y)	−0.213	0.342	0.258	0.244 (0.142–0.716)	<0.001
CRP (mg/L)	0.352	0.615	0.314	2.125 (0.516–4.065)	<0.001
Leukocyte ($10^9/L$)	1.201	1.182	0.870	1.717 (1.310–3.204)	0.001
Erythrocyte ($\times 10^{12}/L$)	2.245	0.706	8.561	8.207 (5.076–11.758)	<0.001
MCH (pg)	0.211	0.659	0.107	1.016 (0.424–2.014)	<0.001
MCHC (g/L)	0.172	0.318	0.304	2.816 (0.921–9.873)	0.031
Hemoglobin (g/L)	1.306	0.457	6.599	3.611 (0.563–8.684)	<0.001
Hematocrit (%)	0.845	0.211	15.163	1.547 (0.127–30.157)	<0.001
Platelet ($10^9/L$)	1.951	0.428	19.281	3.082 (0.055–10.810)	<0.001
PDW (%)	0.366	1.402	0.0507	1.371 (0.643–4.170)	<0.001

CRP: C-reactive protein; MCH: Mean corpuscular hemoglobin; MCHC: Mean corpuscular hemoglobin concentration; PDW: Platelet distribution width.

表4 4种模型的性能评价表

Tab 4 Performance evaluation table of the 4 models

Evaluation index	Logistic	BPNN	LSSVM	Auto-ML
Train data set				
Sensitivity (95%CI)	0.807 (0.553–0.926)	0.924 (0.753–0.981)	0.894 (0.717–0.958)	0.974 (0.780–0.993)
Specificity	0.794	0.915	0.879	0.977
ACC	0.812	0.916	0.885	0.973
AUC (95% CI)	0.816 (0.764–0.881)	0.911 (0.824–0.931)	0.907 (0.765–0.953)	0.988 (0.828–0.997)
Test data set				
Sensitivity (95%CI)	0.716 (0.651–0.892)	0.893 (0.743–0.906)	0.915 (0.792–0.953)	0.967 (0.871–0.994)
Specificity	0.722	0.874	0.815	0.964
ACC	0.822	0.873	0.875	0.973
AUC (95%CI)	0.757 (0.704–0.893)	0.891 (0.812–0.941)	0.908 (0.825–0.949)	0.965 (0.832–0.978)

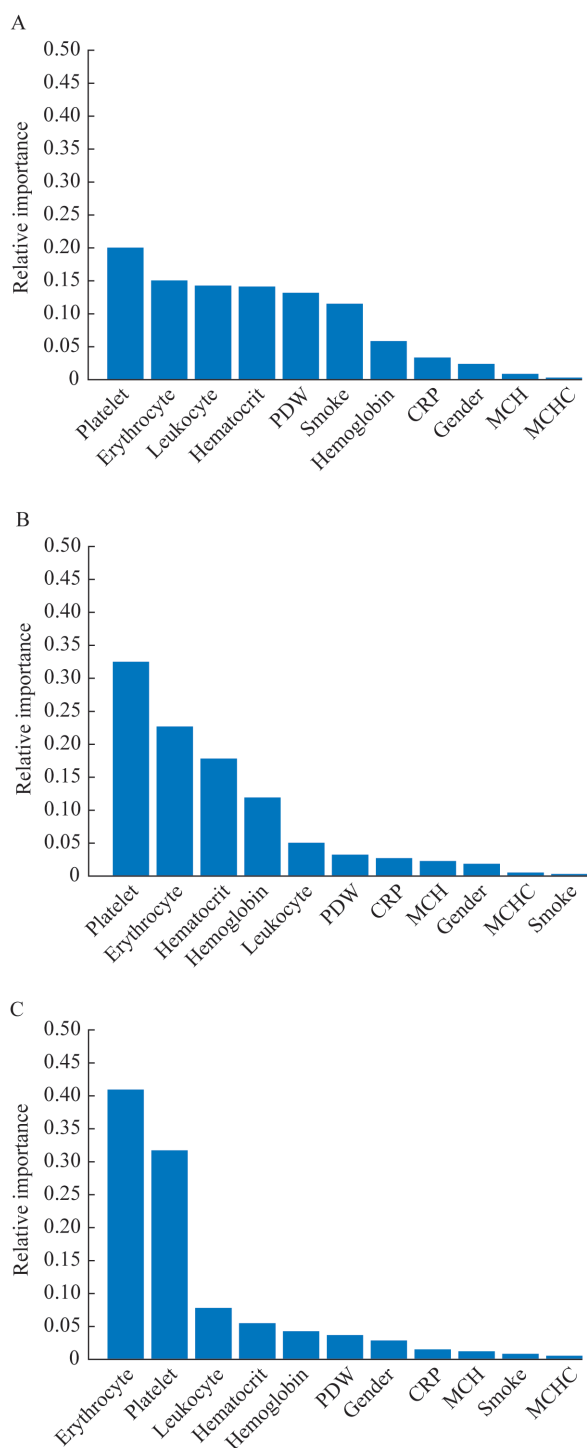
ACC: Accuracy; AUC: Area under the receiver operating characteristic curve.

数,输入2.3节所建立的Auto-ML模型,根据模型输出结果判断每个患者PBSC采集是否成功。将Auto-ML模型判断结果与血液科医师临床诊断结果进行比较,并进一步计算Auto-ML模型的各项评价指标,灵敏度、特异度、准确度、AUC分别为0.984(95%CI: 0.765~0.993)、1.000、0.977、0.981(95%CI: 0.832~0.987),说明Auto-ML模型对PBSC采集结果具有较好的区分度,适合开展临床应用。

讨 论

本研究回顾性分析了在我院血液科行自体PBSC采集术的恶性血液病患者的临床数据。通过多因素Logistic回归分析筛选出影响自体PBSC采

集结果的11项危险因素,包括性别、吸烟史、C反应蛋白、白细胞计数、红细胞计数、平均红细胞血红蛋白含量、平均红细胞血红蛋白浓度、血红蛋白含量、红细胞比积、血小板计数和血小板分布宽度。其中患者C反应蛋白、红细胞计数、血小板计数和血小板分布宽度与既往研究存在差异。有研究表明,自体PBSC采集结果与患者早期化疗时间、化疗程度有关,化疗导致各种血细胞出现不同程度降低^[5,18]。但在细胞刺激因子作用下,PBSC会分化为各种血细胞。PBSC除了具备分化功能,还具备自我增殖能力,且分化和增殖比例通常保持在0.5左右,即化疗后血细胞越多,PBSC越多,而红细胞、血小板在外周血中占比较多,在一定程度上能够反映患者外周血中的PBSC数量。男性患者PBSC采集成功率为84.70%,高于女性患者的72.12%,这与陈晓等^[9]



A: BPNN model indicators; B: LSSVM model indicators; C: Auto-ML model indicators.

图1 3种机器学习模型指标的相对重要性排序

Fig 1 Ranking of the relative importance of three machine learning model indicators

和Mauvais-Jarvis^[19]的研究结果相一致,主要是因为多数血液病患者需要进行不同程度的化疗,造成其免疫能力较低。而女性因雌性激素分泌及月经等

因素而易发生感染,体内免疫系统被唤醒,且在干细胞刺激因子的作用下,CD34⁺和MNC细胞分化为免疫系统功能细胞,因此造成女性体内PBSC少于男性,最终导致女性PBSC采集失败率高于男性。

在模型预测方面,由于Logistic、BPNN、SVM、LSSVM、XGBoost、Auto-ML等机器学习算法区分度高且适应各种非线性问题,因此被广泛用于临床疾病风险预测与诊断,并取得较好的效果。龚军等^[20]和黄浩东等^[21]采用患者一般临床资料、血常规指标、生化指标作等数据,对BPNN、SVM、XGBoost等机器学习算法进行训练,实现了对原发性高血压并发冠心病的患病风险预测,模型精度分别为0.926和0.682。此外,欧笛等^[22]采用SVM、随机森林(Random Forest, RF)、深度学习等算法对乳腺结节良恶性判断进行研究。何文君等^[23]采用RF、Boost、SVM、人工神经网络ANN等机器学习算法构建了AML 1年预后模型,取得了较好的预测效果。然而,在ASCT领域相关研究极度缺乏,尤其是自体PBSC采集结果预测,导致临床医护人员缺少参考。本研究为国内首次采用机器学习模型实现自体PBSC采集结果的预测,并采用5折交叉验证法,分别对Logistic、BPNN、LSSVM、Auto-ML等4种机器学习算法进行训练和验证。Auto-ML模型训练集AUC为0.988(95%CI: 0.828~0.997),验证集为0.965(95%CI: 0.832~0.978),其性能优于其他3种模型,主要是因为Auto-ML具有超强的学习和泛化能力,不需要对模型结构进行调整及其他人工干预,因此在使用过程中能够减少由人工参数调节所带来的模型误差。

虽然Auto-ML模型对自体PBSC采集结果有较好的区分能力,但本研究仍然存在一定的局限性:(1)在指标选取时未纳入患者生化指标的影响。(2)数据来源比较单一,可能存在选择偏倚。(3)虽然采用了部分临床患者数据进行了验证,但是与模型训练集数据相比,数据量仍然偏少,在后续研究中需要收集更多的临床数据对该模型进行验证。

综上,本研究基于我院血液内科就诊的恶性血液病患者的自体PBSC采集数据,筛选出11项PBSC采集结果危险因素,构建了多种机器学习模型,结果显示基于Auto-ML模型对PBSC采集结果具有良好的预测能力。本研究所建立的Auto-ML模型可以应用于临床PBSC采集结果预测和决策支

持系统,从而提高自体PBSC采集的成功率,改善患者生存,减轻患者负担。

作者贡献声明 李若冰 数据统计和分析,论文构思和撰写。唐古生 资助获取,研究设计,论文修订。罗艳蓉 实验操作,制图制表。黄佳莹,张倩倩 实验操作,数据采集和整理。鲁桂华 研究选题和设计,论文指导。

利益冲突声明 所有作者均声明不存在利益冲突。

参 考 文 献

- [1] ANGUITA-COMPAGNON AT, DIBARRART MT, PALMA J, *et al.* Mobilization and collection of peripheral blood stem cells: guidelines for blood volume to process, based on CD34-positive blood cell count in adults and children[J]. *Transplant Proc*, 2010, 42(1):339-344.
- [2] 白敏,王列样,李振华,等.外周血CD34⁺细胞计数对外周血干细胞采集时机选择及结果的影响[J]. *肿瘤研究与临床*, 2021, 33(9):681-684.
- [3] ISHII A, JO T, ARAI Y, *et al.* Development of a quantitative prediction model for peripheral blood stem cell collection yield in the plexixafor era[J]. *Cytotherapy*, 2021, 24(1):49-58.
- [4] 刘忠文,郭建民,杨靖,等.CD34⁺CD38⁻细胞对异基因造血干细胞移植的影响研究[J]. *中国实用内科杂志*, 2011, 31(10):785-768.
- [5] YOSHIKATO T, WATANABE-OKOCHI N, NANNYA Y, *et al.* Prediction model for CD34 positive cell yield in peripheral blood stem cell collection on the fourth day after G-CSF administration in healthy donors[J]. *Int J Hematol*, 2013, 98(1):56-65.
- [6] 程涛.基础血液学[M].北京:科学出版社,2019:79-83.
- [7] FUERST D, HAUBER D, REINHARDT P, *et al.* Gender, cholinesterase, platelet count and red cell count are main predictors of peripheral blood stem cell mobilization in healthy donors[J]. *Vox Sanguinis*, 2019, 114(3):275-282.
- [8] 张红,周芳,宋媛,等.自体造血干细胞移植患者外周血造血干细胞采集影响因素[J]. *白血病·淋巴瘤*, 2019, 28(1):50-51.
- [9] 陈晓,郭智,陈丽娜,等.自体外周血造血干细胞动员采集的影响因素[J]. *中国组织工程研究*, 2021, 25(19):2958-2962.
- [10] LETESTU R, MARZAC C, AUDAT F, *et al.* Use of hematopoietic progenitor cell count on the Sysmex XE-2100 for peripheral blood stem cell harvest monitoring[J]. *Leuk Lymphoma*, 2007, 48(1):89-96.
- [11] NORONHA JFA, LORAND-METZE IGH, GROTTTO HZW. Hematopoietic progenitor cells (HPC) and immature reticulocytes evaluations in mobilization process: new parameters measured by conventional blood cell counter[J]. *J Clin Lab Anal*, 2006, 20(4):149-153.
- [12] TEIPEL R, SCHETELIG J, KRAMER M, *et al.* Prediction of hematopoietic stem cell yield after mobilization with granulocyte-colony-stimulating factor in healthy unrelated donors[J]. *Transfusion*, 2015, 55(12):2855-2863.
- [13] 闫岩,李斯丹,周翊.自体外周血造血干细胞采集时机选择及采集效果预测方法[J]. *中国小儿血液与肿瘤杂志*, 2016, 21(2):104-107.
- [14] 中国抗癌协会淋巴瘤专业委员会,中国医师协会肿瘤医师分会,中国医疗保健国际交流促进会肿瘤内科分会.中国淋巴瘤多学科诊疗模式实施指南[J]. *中华肿瘤杂志*, 2021, 43(2):163-166.
- [15] 中国医师协会血液科医师分会,中华医学会血液学分会,中国医师协会多发性骨髓瘤专业委员会.中国多发性骨髓瘤诊治指南(2020年修订)[J]. *中华内科杂志*, 2020, 59(5):341-346.
- [16] 闰国伟.外周血造血干细胞动员和采集的研究[D].广州:南方医科大学,2014.
- [17] 邹德慧,范磊.造血干细胞移植治疗淋巴瘤中国专家共识(2018版)[J]. *中华肿瘤杂志*, 2018, 40(12):927-934.
- [18] PALMISANO BT, ZHU L, ECKEL RH, *et al.* Sex differences in lipid and lipoprotein metabolism[J]. *Mol Metab*, 2018(15):45-55.
- [19] MAUVAIS-JARVIS F. Gender differences in glucose homeostasis and diabetes[J]. *Physiol Behav*, 2018, 187:20-23.
- [20] 龚军,杜超,钟小钢,等.基于机器学习算法的原发性高血压并发冠心病的患病风险研究[J]. *解放军医学杂志*, 2020, 45(7):735-741.
- [21] 黄浩东,刘小株,龚军,等.基于机器学习算法建立2型糖尿病患者冠心病辅助诊断模型[J]. *复旦学报(医学版)*, 2022, 49(2):226-256.
- [22] 欧笛,姚劲草,李伟,等.基于ABUS和机器学习模型对乳腺结节良恶性判断的对比研究[J]. *肿瘤学杂志*, 2021, 27(12):1001-1005.
- [23] 何文君,石张镇,胡南均,等.构建基于20基因的急性髓系白血病预后生存模型[J]. *中国实验诊断学*, 2021, 25(3):417-420.

(收稿日期:2022-05-13; 编辑:段佳)