Statistical Learning: Final Report

2022 年 10 月 25 日

目录

表格

Models and variables' importance

PTE&DVT vs DVT

表 1: Results of several models to predict PTE&DVT vs DVT

Models	Test Accuracy[1]	Test AUC[2]	Specificity	Sensitivity	Frequency[3]
Logistic-LASSO[4]	0.736(0.015)	0.822(0.012)	0.834(0.018)	0.672(0.027)	0.532(1441)

Note: all missing values have been removed before analysis. 肥胖/高脂暂时因缺失值较多被剔除
[1] Average test accuracy and its standard deviance under 5-fold cross-validation.
[2] Test AUC and its standard deviance under 5-fold cross-validation.
[3] Frequency of PTE, with total number of patients in the bracket.
[4] $\lambda = 0.022$. 非零项 26 个

表 2: Results of several models to predict PTE&DVT vs DVT after splitting age and HR

Models	Test Accuracy[1]	Test AUC[2]	Specificity	Sensitivity	Frequency[3]
Logistic-LASSO[4]	0.734(0.015)	0.821(0.012)	0.831(0.025)	0.670(0.019)	0.532(1441)

Note: all missing values have been removed before analysis. 肥胖/高脂暂时因缺失值较多被剔除
[1] Average test accuracy and its standard deviance under 5-fold cross-validation.
[2] Test AUC and its standard deviance under 5-fold cross-validation.
[3] Frequency of PTE, with total number of patients in the bracket.
[4] $\lambda = 0.023$. 非零项 24 个

Regression Model

表 3: Variables' coefficients from different models

所属类别	Variable	变量名	Model	
			LASSO	p-value[1]
危险因素	sex	性别（男 0，女 1）	0	
	age	年龄	0.002	0.233
	LEF	下肢骨折	0.074	0.41
	HF_Af	心衰或房颤（近 3 月）	-0.079	0.507
	st	严重创伤	0	
	MI	近 3 个月内心肌梗死	0	
	pVTE	既往静脉血栓栓塞	-0.383	0.01*
	SCI	脊髓损伤	0	
	AS	关节镜手术	0	
	AID	自身免疫性疾病	0	
	BT	输血	0	
	CVP	中心静脉插管	0	
	Chemo	肿瘤（转移）/化疗	0	
	NS	肾病综合征	0	
	RS	呼衰	0.014	0.956
	ESA	红细胞生成刺激剂	0.319	0.546
	HRT	激素替代疗法	0	
	Infection	感染	0.172	0.251
	stroke	瘫痪性脑卒中	0	
	SP	表浅静脉血栓形成	-0.382	0.23
	thrombophilia	易栓症	-0.315	0.336
	Postpartum	产后期	0	
	Pregnancy	妊娠	0	
	BS^2	BS^2	0	
	Sedentary	长时间坐位静止不动	-0.482	0.001*

Continued on next page

表 3 – continued from previous page

所属类别	Variable	变量名	Model	
			LASSO	p-value[1]
	hysteroscopy	腹/宫腔镜手术	0	
	Varicose_veins	静脉曲张	0	
	smoking	吸烟	0	
症状	BD	呼吸困难	1.745	0*
	Hemoptysis	咯血	0.243	0.303
	chest_pain	胸痛	0	
	LLS	下肢肿痛	-0.266	0.049*
	ULS	上肢肿痛	0	
	fever	发热	0	
	Syncope	头晕/晕厥	0.467	0.164
	dry_cough	干咳或伴有喘息	0	
	Palpitation	心慌/心悸	0	
	irritability	烦躁/谵妄/意识障碍	0	
	AM	皮肤或肢端湿冷	0.089	0.688
	LC	口唇青紫	0	
体征	tachycardia	心动过速	0.25	0.049*
	DR	呼吸音减弱	0.586	0.051
	Lung_rales	肺部啰音	0.233	0.317
	P2Hyper	P2 亢进或分裂	0.225	0.036*
	TVSM	三尖瓣收缩期杂音	0	
	HJV[3]	HJV[3]	0	
	GMT	腓肠肌压痛，活动受限	0.088	0.112
心电图	HR	心率（次/分）	0	
	S1Q2T3	SIQ2T3	0.768	0*
	ST	严重创伤	0	
	AT	房速	0	
	RVH	右心室肥大	0	

Continued on next page

表 3 – continued from previous page

所属类别	Variable	变量名	Model	
			LASSO	p-value[1]
	p_pulmonale	肺性 P 波	0	
	CEARD	电轴右偏	0	
	CEALD	电轴左偏	0.726	0*
	S1S2S3	S1S2S3	0.82	0.004*
	low_voltage	低电压	0	
	clockwise	顺时针旋转	0	
	ST_SE	ST 段抬高	0	
	ST_SD	ST 段压低	0	
	T_V13_4	T 波倒置（V1-V3/V4）	0.358	0.05*
	ST2_3_AVF	ST 抬高 II/III/AVF	0	
	STD2_3_AVF	ST 压低 II/III/AVF	0.096	0.588
	q_Q2_AVF	q/Q II/AVF	0.212	0.206
	T_2_AVF	T 波倒置 II/AVF	0	
	RBBB	右束支传导阻滞	0	

[1] Logistic-LASSO 的 p 值 (Lee et al. 2016)
[2] BS: 卧床 >3 天或接受外科手术
[3] HJV: 颈静脉充盈或搏动，肝颈静脉回流征阳性

表 3 展示的是用剔除缺失值较多的变量“肥胖/高脂”，并移除少数缺失值后用所有变量做 Logistic 回归的模型系数及其 P 值。Logistic 回归中，采用 LASSO 筛选出的非零项为 26 个。

表 4: List of significant variables in a Logistic regression model

Type	Variable	变量名	Coef.	OR	P-value
危险因素	pVTE	既往静脉血栓栓塞	-0.383	0.682	0.01*
	Sedentary	长时间坐位静止不动	-0.482	0.618	0.001*
症状	BD	呼吸困难	1.745	5.726	0*
	LLS	下肢肿痛	-0.266	0.766	0.049*
体征	tachycardia	心动过速	0.25	1.284	0.049*
	P2Hyper	P2 亢进或分裂	0.225	1.252	0.036*
心电图	S1Q2T3	SIQ2T3	0.768	2.155	0*
	CEALD	电轴左偏	0.726	2.067	0*
	S1S2S3	S1S2S3	0.82	2.270	0.004*
	T_V13_4	T 波倒置（V1-V3/V4）	0.358	1.430	0.05*

Logistic-LASSO 中显著的预测变量，其系数及所属类别如上表（表 4）所示。

而将年龄与心率分组之后的模型结果如下表（表 5）所示。

表 5: Variables' coefficients from different models after splitting

所属类别	Variable	变量名	Model	
			LASSO	p-value[1]
	sex	性别（男 0，女 1）	0	
年龄	age_group1	[10, 20)	baseline	
	age_group2	[20, 30)	0	
	age_group3	[30, 40)	0	
	age_group4	[40, 50)	0	
	age_group5	[50, 60)	0	
	age_group6	[60, 70)	0	
	age_group7	[70, 80)	0	
	age_group8	[80, 90)	0	
	age_group9	[90, 100)	0	
危险因素	LEF	下肢骨折	0.065	0.463
	HF_Af	心衰或房颤（近 3 月）	-0.008	0.946
	st	严重创伤	0	
	MI	近 3 个月内心肌梗死	0	
	pVTE	既往静脉血栓栓塞	-0.373	0.096
	SCI	脊髓损伤	0	
	AS	关节镜手术	0	
	AID	自身免疫性疾病	0	
	BT	输血	0	
	CVP	中心静脉插管	0	
	Chemo	肿瘤（转移）/化疗	0	
	NS	肾病综合征	0	
	RS	呼衰	0	
	ESA	红细胞生成刺激剂	0.174	0.73
	HRT	激素替代疗法	0	

Continued on next page

表 5 – continued from previous page

所属类别	Variable	变量名	Model	
			LASSO	p-value[1]
	Infection	感染	0.161	0.449
	stroke	瘫痪性脑卒中	0	
	SP	表浅静脉血栓形成	-0.339	0.15
	thrombophilia	易栓症	-0.273	0.371
	Postpartum	产后期	0	
	Pregnancy	妊娠	0	
	BS^2	BS^2	0	
	Sedentary	长时间坐位静止不动	-0.465	0.026*
	hysteroscopy	腹/宫腔镜手术	0	
	Varicose_veins	静脉曲张	0	
	smoking	吸烟	0	
症状	BD	呼吸困难	1.737	0*
	Hemoptysis	咯血	0.196	0.332
	chest_pain	胸痛	0	
	LLS	下肢肿痛	-0.256	0.552
	ULS	上肢肿痛	0	
	fever	发热	0	
	Syncope	头晕/晕厥	0.448	0.02*
	dry_cough	干咳或伴有喘息	0	
	Palpitation	心慌/心悸	0	
	irritability	烦躁/谵妄/意识障碍	0	
	AM	皮肤或肢端湿冷	0.075	0.709
	LC	口唇青紫	0	
体征	tachycardia	心动过速	0.233	0.166
	DR	呼吸音减弱	0.561	0.052
	Lung_rales	肺部啰音	0.234	0.251
	P2Hyper	P2 亢进或分裂	0.218	0.149

Continued on next page

表 5 – continued from previous page

所属类别	Variable	变量名	Model	
			LASSO	p-value[1]
	TVSM	三尖瓣收缩期杂音	0	
	HJV[3]	HJV[3]	0	
	GMT	腓肠肌压痛，活动受限	0.071	0.171
心电图	HR_group1	心率 <60	baseline	
	HR_group2	心率 [60, 100)	0	
	HR_group3	心率 ≥100	0	
	S1Q2T3	SIQ2T3	0.748	0.281
	ST	严重创伤	0	
	AT	房速	0	
	RVH	右心室肥大	0	
	p_pulmonale	肺性 P 波	0	
	CEARD	电轴右偏	0	
	CEALD	电轴左偏	0.727	0*
	S1S2S3	S1S2S3	0.782	0.005*
	low_voltage	低电压	0	
	clockwise	顺时针旋转	0	
	ST_SE	ST 段抬高	0	
	ST_SD	ST 段压低	0	
	T_V13_4	T 波倒置（V1-V3/V4）	0.355	0.047*
	ST2_3_AVF	ST 抬高 II/III/AVF	0	
	STD2_3_AVF	ST 压低 II/III/AVF	0.09	0.614
	q_Q2_AVF	q/Q II/AVF	0.2	0.289
	T_2_AVF	T 波倒置 II/AVF	0	
	RBBB	右束支传导阻滞	0	

[1] Logistic-LASSO 的 p 值 (Lee et al. 2016)
[2] BS: 卧床 >3 天或接受外科手术
[3] HJV: 颈静脉充盈或搏动，肝颈静脉回流征阳性

表 6: List of significant variables in a Logistic regression model after splitting

Type	Variable	变量名	Coef.	OR	P-value
危险因素	Sedentary	长时间坐位静止不动	-0.465	0.628	0.026*
症状	BD	呼吸困难	1.737	5.680	0*
	Syncope	头晕/晕厥	0.448	1.565	0.02*
心电图	CEALD	电轴左偏	0.727	2.069	0*
	S1S2S3	S1S2S3	0.782	2.186	0.005*
	T_V13_4	T 波倒置（V1-V3/V4）	0.355	1.426	0.047*

将年龄和心率分组后 Logistic-LASSO 中显著的预测变量，其系数及所属类别如上表（表 6）所示。