第 1 章
临床预测模型基础

本书认为统计可以分三个维度：即初级统计说一说——描述性分析（指标、图表）、中级统计比一比——差异性分析（假设检验）、高级统计找关系——关系性分析（统计模型）。高级统计找关系，也就是构建统计学模型，是统计学的集大成，是统计探索复杂事物背后规律的至高境界。

临床预测模型属于高级统计找关系，因此，学习需要一定初级和中级的基础，同时相对于常规模型而言，临床预测模型更加注重构建模型后的验证和评价，其有自己独特的建模策略和评价指标体系。本章将为大家做一个系统的介绍。

1.1 三种建模策略解读

生物医药领域构建模型，根据研究目的的不同，可以有三种模型：①风险因素发现模型，其目的是探索某疾病或某健康事件发生的可能风险因素，研究重点放在可能的多个影响因素 X；②风险因素验证模型，其目的是验证某因素在某疾病或健康事件发生中的作用大小，研究重点放在某个需要证明其作用大小的某个 X；③临床预测模型，其目的是通过构建一个多因素模型，预测某疾病发生或某预后结局发生的概率，研究重点放在 Y，放在模型的预测效果，希望构建一个预测更加准确有效的模型。三种模型的目的不同，因此，构建模型时的建模策略也不尽相同。

1.1.1 风险因素发现模型

风险因素发现模型常采用的是“先单后多策略”，即对专业上认为可能有意义的风险因素进行单因素分析，对单因素分析有统计学意义的风险因素，再进行多因素分析，最终以 $P < 0.05$ 为判断标准，确定某疾病最终的风险因素。先单后多有两种方法：一是单因素分析采用差异性分析的假设检验，如图 1-1 所示，差异性分析发现 7 个可能的风险因素，然后将 7 个可能的风险因素一起纳入模型进行多因素分析 PK，如图 1-2 所示；二是直接采用单因素回归分析，选择有意义的因素，再一起纳入多因素回归分析，如图 1-3 所示。国内文章以第一种居多，近年有向第二种转变的趋势，而 SCI 论文往往以第二种居多。

项目		例数	口渴感评分（$\bar{x} \pm s$）	t/F	P
性别	男	69	6.18 ± 1.02	0.143	0.887
	女	39	6.11 ± 1.09		
年龄（岁）	47 ~	67	6.17 ± 1.40	0.107	0.915
	61 ~ 88	41	6.10 ± 1.42		
疾病	呼吸系统	51	6.19 ± 1.33	0.028	0.994
	循环系统	27	6.13 ± 1.29		
	神经系统	19	6.10 ± 1.32		
	其他	11	6.18 ± 1.31		
APACHE-Ⅱ	0 ~	61	6.15 ± 1.05	0.036	0.965
评分（分）	10 ~	39	6.09 ± 1.14		
	20 ~ 26	8	6.12 ± 1.13		
入住 ICU 天数	1 ~	22	6.12 ± 1.23	0.047	0.954
（d）	3 ~	45	6.14 ± 1.28		
	7 ~ 23	41	6.21 ± 1.30		
呼吸衰竭	Ⅰ型	64	6.16 ± 1.24	0.164	0.871
类型	Ⅱ型	44	6.12 ± 1.26		
无创机械	首次	85	7.08 ± 1.67	5.716	0.000
通气	非首次	23	4.95 ± 1.21		
机械通气	0 ~	108	5.59 ± 1.13	43.666	0.000
时间（h）	4 ~	108	6.78 ± 1.39		
	8 ~ 12	108	7.21 ± 1.42		
呼吸机类型	双水平	45	6.12 ± 1.21	0.124	0.902
	非双水平	63	6.18 ± 1.19		
呼吸型态	张口	52	7.09 ± 1.24	5.131	0.000
	闭口	56	5.89 ± 1.19		
吸入潮气量	6.2 ~	23	5.98 ± 1.24	4.474	0.014
（mL/kg）	8.0 ~	55	6.42 ± 1.26		
	10.0 ~ 16.7	30	7.01 ± 1.30		
呼吸频率	8 ~	13	5.99 ± 1.02	10.530	0.000
（次/min）	12 ~	41	6.59 ± 1.78		
	20 ~ 39	54	6.97 ± 1.90		
湿化温度	32 ~	49	6.10 ± 1.42	1.499	0.137
（℃）	35 ~ 37	59	6.51 ± 1.41		
漏气量	9 ~	41	5.88 ± 1.31	4.954	0.009
（mL/次）	50 ~	32	6.31 ± 1.56		
	100 ~ 348	35	6.98 ± 1.71		
面罩舒适度	好	51	4.95 ± 1.05	31.958	0.000
	一般	42	6.45 ± 1.14		
	差	15	7.02 ± 1.13		

图 1-1　单因素分析（差异性检验）

变量	β	SE	$Wald\chi^2$	P	OR	95% CI
是否首次使用	−2.196	1.095	4.019	0.045	0.111	0.013 ~ 0.952
呼吸型态	2.185	1.027	4.526	0.033	8.891	1.188 ~ 66.551
机械通气时间	1.470	0.604	5.920	0.015	4.350	1.331 ~ 14.218
面罩舒适度	1.245	0.576	4.661	0.031	3.472	1.122 ~ 10.746
漏气量	1.428	0.626	5.206	0.023	4.169	1.223 ~ 14.214
呼吸频率	1.291	0.575	5.043	0.025	3.636	1.178 ~ 11.220
吸入潮气量	1.264	0.614	4.223	0.040	3.539	1.062 ~ 11.797
常量	−12.311	3.696	11.096	0.001	–	–

注：自变量赋值，是否首次使用 1 = 是，2 = 否；呼吸型态 1 = 张口，2 = 闭口；机械通气时间 1 = 0 ~ h，2 = 4 ~ h，3 = 8 ~ 12 h；面罩舒适度 1 = 好，2 = 一般，3 = 差；漏气量 1 = 9 ~ mL/次，2 = 50 ~ mL/次，3 = 100 ~ 348 mL/次；呼吸频率 1 = 8 ~ 次/min，2 = 12 ~ 次/min，3 = 20 ~ 39 次/min；吸入潮气量 1 = 6.2 ~ mL/kg，2 = 8.0 ~ mL/kg，3 = 10.0 ~ 16.7 mL/kg。

图 1-2　差异性分析后多因素分析

影响因素	B	S.E.	Wals	P	Exp(B)	95%CI for Exp(B)	
						低	高
年龄	−0.125	0.263	0.224	0.636	0.883	0.527	1.479
手术时间	−0.123	0.385	0.101	0.750	0.885	0.416	1.883
术中出血量	0.698	0.282	6.149	0.013	2.010	1.158	3.491
是否合并糖尿病	1.897	0.664	8.164	0.004	6.667	1.184	24.494
是否合并高血压	0.531	0.497	1.140	0.286	1.701	0.642	4.508
是否合并心脏病	0.769	1.240	0.384	0.535	0.464	0.041	5.269
是否合并肾功能不全	1.713	0.801	4.573	0.032	5.543	1.154	26.635
是否合并 COPD	−0.058	0.840	0.005	0.945	0.943	0.182	4.894
术后血红蛋白水平	−0.912	0.265	11.825	0.001	0.402	0.239	0.676
术后白蛋白水平	−0.873	0.246	12.56	0.000	0.418	0.258	0.677
肿瘤分期	0.473	0.388	1.488	0.222	1.605	0.751	3.432

影响因素	B	S.E.	Wals	P	Exp(B)	95%CI for Exp(B)	
						低	高
术中出血量	0.864	0.338	6.550	0.010	2.373	1.224	4.599
是否合并糖尿病	1.704	0.790	4.650	0.031	5.495	1.168	25.855
是否合并肾功能不全	2.248	0.994	5.109	0.024	9.467	1.348	66.483
术后血红蛋白水平	−0.689	0.285	5.854	0.016	0.502	0.287	0.877
术后白蛋白水平	−0.715	0.307	5.432	0.020	0.489	0.268	0.893

图 1-3　单因素回归与多因素回归分析

1.1.2　风险因素验证模型

验证风险模型建模常采用“抽丝剥茧策略”或“层层加码策略”。其建模首先进行该因素 X 与 Y 的单因素回归，其次逐渐往模型中添加不同的控制因素，以期发现在控制了若干可能的混杂因素之后，最后证明该因素 X 是否是 Y 的风险因素，并确定其风险大小。

如图 1-4 所示，该作者为了验证 serum sphingomyelin 与 CHD 发病的关系，先构建了 Model1，单独研究 serum sphingomyelin 与 CHD 发病关系得到 HR（常称为 crude HR 或 unadjusted HR），结果发现 serum sphingomyelin 每改变 1 个单位，CHD 发生风险增加 44%（$P < 0.001$）；然后在 Model1 的基础上构建 Model2，在 serum sphingomyelin 基础上，模型增加了性别、糖尿病发病年龄、糖尿病病程与吸烟，结果发现 serum sphingomyelin 依旧与 CHD 发

病有关，HR=1.24（*P*=0.038）；然后继续在 Model2 基础上增加变量构建 Model3，如此反复直至 Model8，最终在 Model8 中发现，serum sphingomyelin 与 CHD 发病风险并无关系，虽然 HR=1.16，但 *P*=0.18 已经无统计学意义，所以最终证明 serum sphingomyelin 可以“无罪释放”。

Model	Incident CHD	
	HR[a] (95% CI)	*p* value
Model 1: Serum sphingomyelin	1.44 (1.19, 1.73)	<0.001
Model 2: Model 1 + sex + age of diabetes onset + diabetes duration + smoking	1.24 (1.01, 1.52)	0.038
Model 3: Model 2 + systolic blood pressure	1.21 (0.98, 1.49)	0.07
Model 4: Model 3 + HDL-cholesterol	1.21 (0.99, 1.48)	0.06
Model 5: Model 4 + triacylglycerols	1.17 (0.96, 1.43)	0.13
Model 6: Model 5 + BMI	1.18 (0.96, 1.44)	0.11
Model 7: Model 6 + HbA_{1c}	1.13 (0.92, 1.39)	0.24
Model 8: Model 3 + BMI + HbA_{1c}	1.16 (0.94, 1.43)	0.18

The incidence of CHD was 8.2 per 1000 person-years

[a] Reported HR increase corresponds to a 1 SD increase in sphingomyelin level

图 1-4　serum sphingomyelin 与 CHD 发病风险验证

通过这种在要验证的核心 *X* 基础上“逐层加码”，对 *X* 与 *Y* 之间的关系进行“抽丝剥茧”，最终对 *X* 是否是导致 *Y* 发生的风险因素进行验证，从某种程度上，松哥认为比风险因素发现模型要更具价值，因为这类模型可以对某 *X* 进行“最终审判”；图 1-5 也展示了这种验证风险模型的构建策略。

	B	Wald	*P*	OR	95%CI	
					下限	上限
Model1	0.038	5.607	0.018	1.039	1.007	1.072
Model2	0.040	5.971	0.015	1.040	1.008	1.074
Model3	0.043	6.691	0.010	1.044	1.011	1.079
Model4	0.046	5.593	0.018	1.047	1.008	1.088
Model5	0.274	0.818	0.366	1.315	0.726	2.384

注：Model 1：校正前

Model 2：校正 Sex、Age

Model 3：校正 Sex、Age、BMI、WHR

Model 4：校正 Sex、Age、BMI、WHR、SBP、DBP

Model 5：校正 Sex、Age、BMI、WHR、SBP、DBP、TG、TC、HDL-C、LDL

图 1-5　验证性风险因素建模展示

1.1.3　临床预测模型

临床预测模型是指利用多因素模型估算患有某病的概率或将来某结局的发生概率，主要分为诊断模型（diagnostic model）和预后模型（prognostic model）。

诊断模型主要基于研究对象的临床特征，预测当前患有某种疾病的概率，多见于横断面研究，病例对照研究；预后模型则是针对患有某种疾病的研究对象，预测将来疾病复发、

死亡、伤残等转归的概率，多见于纵向研究。

临床预测模型建模策略依旧采用的是“先单后多策略”，但是其重点在于对 Y 预测的准确性，即不再对模型中每一个 X 是否 $P < 0.05$ 纠结，只要模型整体预测效果好，可以包容 $P > 0.05$ 的 X 在模型内的存在，此时模型优劣判定往往按照 AIC 准则进行，图 1-6 和图 1-7 反映了预测模型先单后多的建模策略。

Identification of predictive factors at hospital admission for pharmacist interventions with a clinical impact. Laboratory results, clinical characteristics and medication data were analyzed with univariate and multivariable regression.

	Univariate model			Multivariate model		
Variable	**OR**	**95%CI**	***P*-value**	**OR**	**95%CI**	***P*-value**
Nervous system drug class on BPMH	2.33	1.17–4.64	0.02			ns
Cardiovascular drug class on BPMH	2.49	1.20–5.14	0.01			ns
Number of medication on BPMH ⩾ 5	4.67	2.01–10.83	<0.001	3.03	1.29–7.51	0.01
Charlson Comorbidity index score ⩾ 2	4.36	2.12–8.96	<0.001	3.13	1.49–6.71	<0.01

ATC = anatomic therapeutic chemical, BPMH = Best Possible Medication History, CI = confidence interval, ns = nonsignificant, OR = odds ratio,.

图 1-6　临床预测模型先单后多 Logistic 回归建模策略展示

	Univariate analysis			Multivariate analysis		
	HR	**95%CI**	***P***	**HR**	**95%CI**	***P***
Gender (female vs. male)	1.099	0.565–2.137	0.782			
Age (⩾69 vs. <69; y)	1.997	0.905–4.408	0.087	3.253	1.353–7.825	0.008
Tumor size (⩾4.2 vs. <4.2; cm)	2.945	1.376–6.304	0.005			
TNM Stage (III–IV vs. I–II)	4.396	2.230–8.664	0.000	4.712	2.272–9.770	0.000
IgG (⩾12.51 vs. <12.51; g/L)	0.457	0.225–0.926	0.030	0.385	0.176–0.840	0.017
IgA (⩾1.97 vs. <1.97; g/L)	1.474	0.773–2.812	0.239			
IgM (⩾1.09 vs. <1.09; g/L)	0.538	0.273–1.060	0.073			
C3 (⩾1.05 vs. <1.05; g/L)	1.924	0.967–3.830	0.062			
C4 (⩾0.22 vs. <0.22; g/L)	1.582	0.773–3.235	0.209			
BF (⩾0.47 vs. <0.47; g/L)	2.087	0.876–4.968	0.097			
CRP (⩾3.78 vs. <3.78; mg/L)	1.621	0.722–3.638	0.241			
WBC(⩾4.95 vs. <4.95; 10^9/L)	0.467	0.218–1.003	0.051			
LMR (⩾4.15 vs. <4.15)	0.243	0.110–0.539	0.001	0.408	0.177–0.940	0.035
PLR (⩾177.86 vs. <177.86)	2.205	0.913–5.325	0.079			
NLR (⩾1.61 vs. <1.61)	0.682	0.351–1.328	0.261			

HR, Hazard ratio; 95% CI, 95% confidence interval; TNM, tumor/node/metastasis; IgG, immunoglobulin G; IgA, immunoglobulin A; IgM, immunoglobulin M; C3, Complement 3; C4, Complement 4; BF, B factor; CRP, C-reactive protein; WBC, white blood cell count; LMR, lymphocyte-to-monocyte ratio; PLR, platelet-to-lymphocyte ratio; NLR, neutrophil-to-lymphocyte ratio; OS, overall survival

图 1-7　临床预测模型先单后多 COX 回归建模策略展示

临床预测模型的建模，有一个“门当户对”原则，这个虽然统计教材中没有说，但确实是数据处理的经验累积。“门当户对”是指，我们研究的变量从性质而论，有定量与定性两类，建模时尽量满足因变量与自变量的定量对定量，定性对定性。这就是“门当户对”原则。

临床预测模型中 Logistic 回归的因变量为二分类变量（注意临床预测模型中的 Logistic 回归只是 Binary Logistic regression，其他多项和有序资料的临床预测模型方法尚不成熟）；COX 回归的因变量是二分类 + 时间；所以构建模型时的自变量（风险因素或预测因子）如果是定性则会较好，因为满足“门当户对”的原则，如果 Logistic 或 COX 回归中，纳入的是定量变量，也许统计上有意义，也能解释，但是专业上可能不太容易解释。

比如说年龄，如果直接代入，那结果解释则为年龄每增加 1 岁，发生某种疾病或结局的风险增加多少，统计上没问题，但是试问大家，哪种疾病只要增加 1 岁，就会增加专业

上有意义的风险呢？所以，为什么大家经常看到，很多文章会把年龄进行分组，如小于 60 岁和大于等于 60 岁等。

故而，您再看上面的图 1-6 和图 1-7，其中的那么多原本是定量的指标，均根据专业进行了变量降维，从定量降维为定性，如图 1-7 中的年龄，分为大于等于 69 岁和小于 69 岁。看到这您也许会心存困惑，为什么年龄降维分组，在不同文献中往往不一样呢？是的，年龄的降维分组，文献中不下 10 种方法，没有固定的套路，需要您根据自己的专业或者数据的特征进行降维，具体的请看本书相关章节。

1.2　临床预测模型分类与分型

根据研究目的的不同、建模与验证数据集的不同，临床预测模型可以有不同的分类。

1.2.1　预测模型目的分类

如果预测模型是为了预测研究对象是否患有某种疾病，采用的是 Logistic 回归建模，数据来源于横断面的研究，我们称之为诊断预测模型，如图 1-8 所示；如果我们建模是为了预测某病患者将来某种预后结局的发生风险，采用的是 COX 回归建模，数据来源于队列研究，我们称之为预后预测模型，如图 1-9 所示。

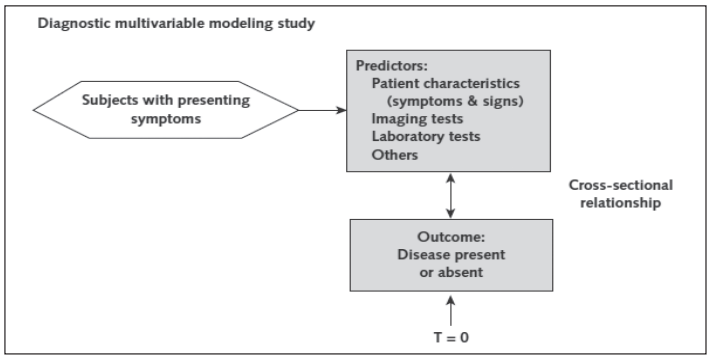

图 1-8　诊断预测模型模式图

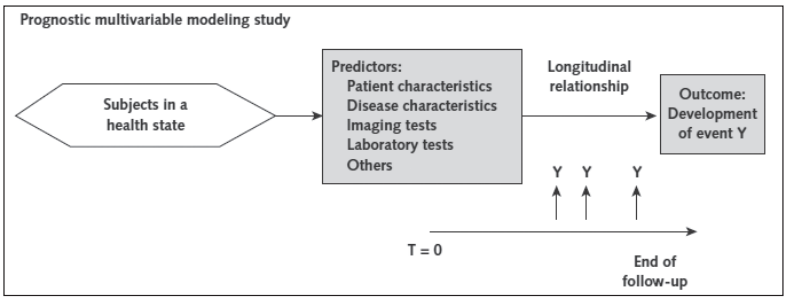

图 1-9　预后预测模型模式图

1.2.2 预测模型数据来源分类

临床预测模型区别于普通模型的要点之一在于验证，就是将构建的模型代入某数据集进行预测，以评价其表现。根据数据集的不同，预测模型又有如下不同的分型，如图 1-10 和图 1-11 所示。

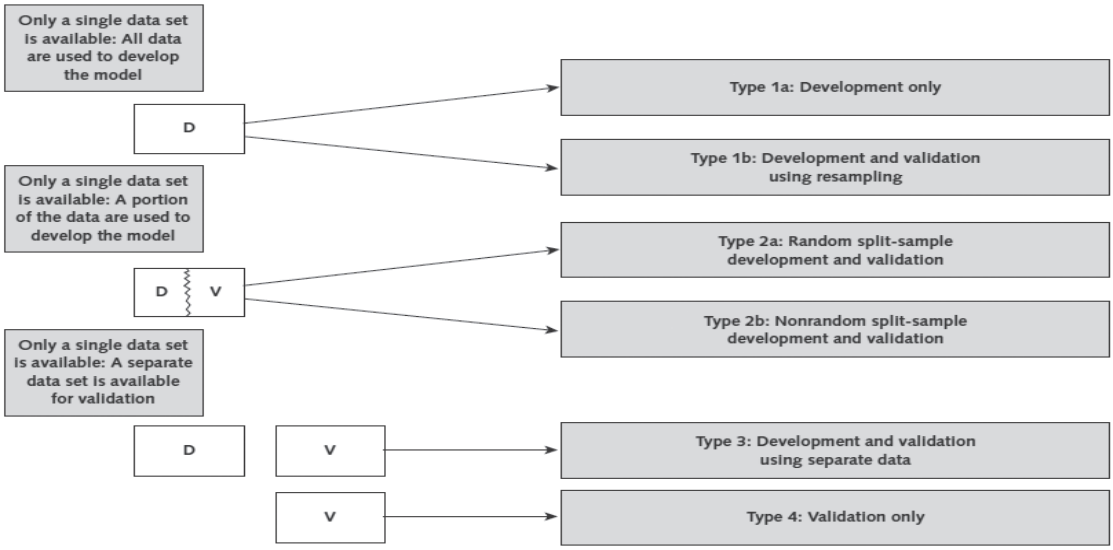

图 1-10 预测模型按照数据集分型图解

Analysis Type	Description
Type 1a	Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance).
Type 1b	Development of a prediction model using the entire data set, but then using resampling (e.g., bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model. Resampling techniques, generally referred to as "internal validation", are recommended as a prerequisite for prediction model development, particularly if data are limited (6, 14, 15).
Type 2a	The data are randomly split into 2 groups: one to develop the prediction model, and one to evaluate its predictive performance. This design is generally not recommended or better than type 1b, particularly in case of limited data, because it leads to lack of power during model development and validation (14, 15, 16).
Type 2b	The data are nonrandomly split (e.g., by location or time) into 2 groups: one to develop the prediction model and one to evaluate its predictive performance. Type 2b is a stronger design for evaluating model performance than type 2a, because allows for nonrandom variation between the 2 data sets (6, 13, 17).
Type 3	Development of a prediction model using 1 data set and an evaluation of its performance on separate data (e.g., from a different study).
Type 4	The evaluation of the predictive performance of an existing (published) prediction model on separate data (13).
Types 3 and 4 are commonly referred to as "external validation studies." Arguably type 2b is as well, although it may be considered an intermediary between internal and external validation.	

D = development data; V = validation data.

图 1-11 各型预测模型解读

Type1a：仅仅在训练集中构建临床预测模型。大家是否发现，这与我们前面所谓的风险因素发现模型高度相似？

Type1b：在训练集中进行建模，然后利用重抽样的方法，在训练集中抽取样本进行验证。

Type2a：将研究数据集随机拆分为训练集和验证集，然后用训练集建模，用验证集进行验证。

Type2b：将研究数据集进行非随机拆分为训练集和验证集，然后用训练集建模，用验证集进行验证。如在某医院，搜集前 2 年的数据作为训练集，搜集后 1 年的数据用于验证。

Type3：用不同的数据集进行建模和验证，如用某医院的数据建模，用另外一家医院的数据进行验证。

Type4：仅进行验证。如别人已经构建了一个临床预测模型并且发表了文章，你用该模型在自家医院进行验证。

另从撰写角度，围绕某个临床主题，预测模型撰写的过程如下：①人无我有，即目前尚无临床预测模型，新建一个用于预测；②人有我优，即别人发表了一个临床预测模型，你可以对该模型进行改良和优化；③人有我验，即对别人发表的预测模型，你可以用自己的数据进行验证；④人多我比，即已经有多个临床预测模型，你可以用自己的数据，对多个模型的表现进行比较。

1.2.3　数据集分类

临床预测模型建模和验证的数据集称为训练集和验证集。如果验证集是来自训练集来源的同一人群，则称为内部验证；如果验证集来自外部数据，则称为外部验证，如图 1-12 所示。

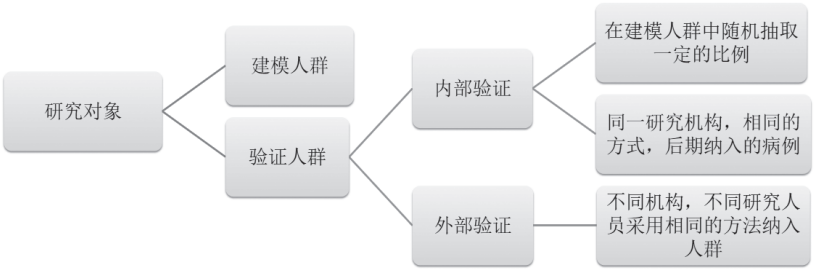

图 1-12　预测模型数据集与验证区分

很多时候，我们对数据进行拆分，分为训练集和验证集，用训练集建模，然后用验证集进行验证，注意此时也是内部验证，因为验证集与训练集是同一批数据被拆分开的；同样在某单位搜集前 2 年的数据作为训练集，后 1 年的数据作为验证集，这种情况也属于内部验证，因为也是来自同一家机构的。外部验证是指验证集来自不同于训练集的数据，比如 A 医院建模，而采用 B 医院数据进行验证，这就属于外部验证。

另从统计分析方法上，同一数据进行拆分为训练集和验证集，与 A 医院数据建模，B 医院数据验证，在统计实现方法上是一样的，均属于统计学上的外部验证；而对研究全部数据进行建模，然后采用 Bootsrap 或交叉验证的方法验证，都属于统计学上的内部验证。这个容易混淆，一个是统计学上的内部验证与外部验证，另一个是数据集上的内部验证与外部验证。

1.3 区分度 -C 指数

区分度（Discrimination）是指预测模型把发生与未发生某结局事件的受试对象区分开来的能力。简单来说就是把患者与非患者区分开的能力，反映模型是否能“明辨是非”。区分度是模型对结局事件的一种定性判断。

区分度评价常用指标包括：AUC/C-index、NRI、IDI（NRI 和 IDI 用于新旧模型区分度的比较），对于诊断预测模型和预后预测模型也稍有区别。诊断预测模型 Logistic 回归指标包括：AUC、ROC、NRI、IDI；预后预测模型 COX 回归区分度指标包括：C-index、Time-ROC、NRI、IDI，Time_AUC，其中对于预后模型，C-index 是针对整个模型的，其他都可以分时间点，如 1year、3year 和 5year ROC，以及对应时点的 NRI 和 IDI。Time_AUC 是将要研究的各个时间点的 Time-ROC 曲线下面积绘制一条曲线，可以综合反映模型的总体区分度的情况。本节我们先学习 C-index。

C-index 常写为 Harrell concordance index、C statistics、C-indices、Concordance indices，对于 Logistic 回归，C-index 就是 ROC 分析的 AUC（Area Under Curve）。C 指数的一般判定标准如下。

- C-index：0.5　　完全不一致
- C-index：1.0　　完全一致
- C-index：0.5-0.7　　较低区分度
- C-index：0.71-0.90　　中等区分度
- C-index：＞ 0.90　　高度区分度

图 1-13 为诊断预测模型区分度 ROC 曲线，其中左为训练集，右为验证集，每个 ROC 曲线中又包含 3 条 ROC 曲线，那是因为构建了 3 个模型，将 3 个模型的 ROC 曲线进行了合并绘制。

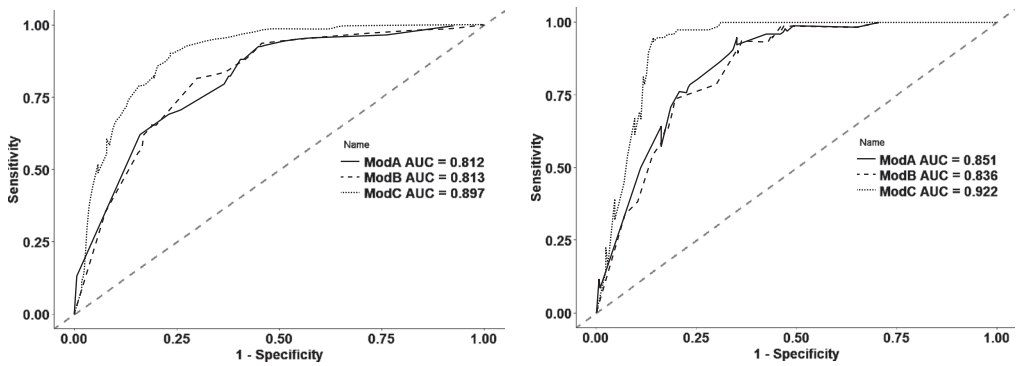

图 1-13　诊断模型的 ROC 分析

对于预后模型，区分度评价可以计算模型整体的 C-index 以及绘制时点 ROC 曲线，如图 1-14 和图 1-15 所示。图 1-14 左图和图 1-14 右图分别为训练集和验证集 6 个月、12 个月

和 24 个月 ROC 曲线，并且给出各自的 AUC。

图 1-15 上面 3 幅为训练集 1、3、5 年时点 ROC，图 1-15 下面 3 幅为验证集 1、3、5 年时点 ROC，图中的两条 ROC 曲线是作者构建的 Nomogram 模型和 TNM stage 模型，以及各自模型的 AUC。

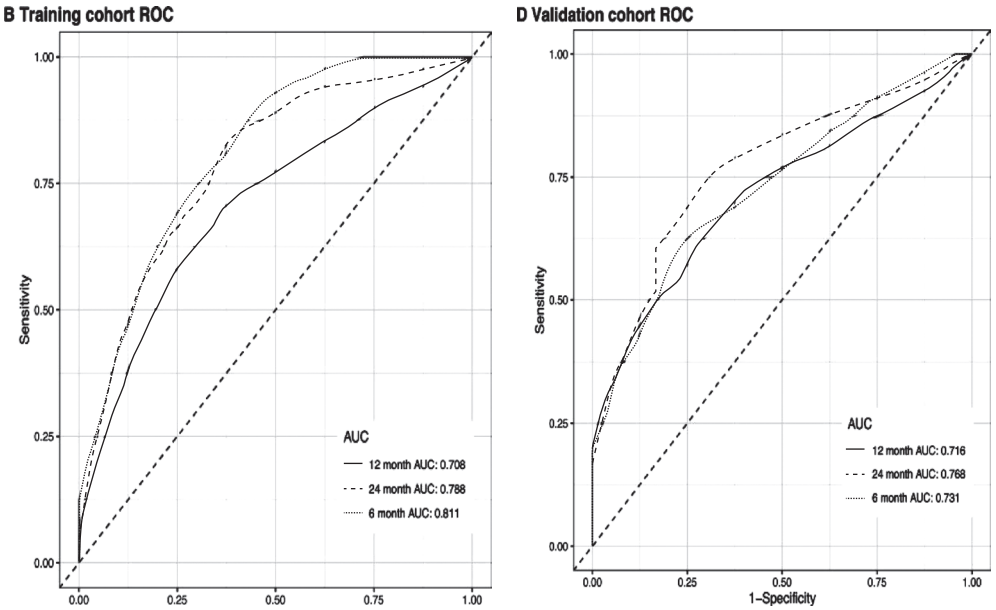

图 1-14　预后模型时点依赖 ROC 曲线方式（一）

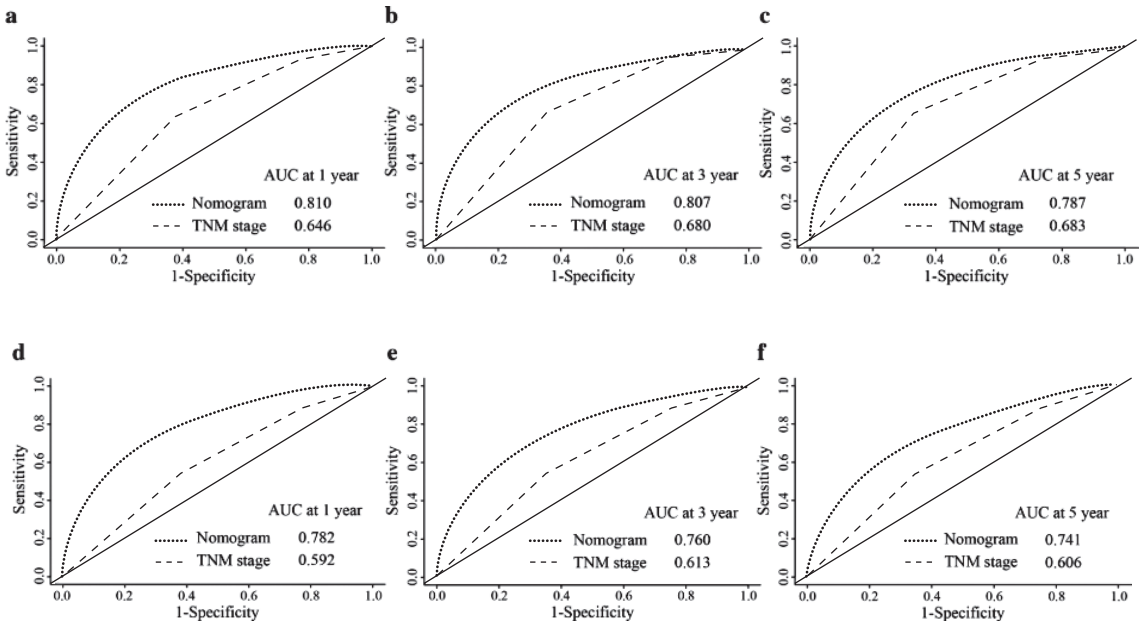

图 1-15　预后模型时点依赖 ROC 曲线方式（二）

上面仅展示了诊断与预后模型共同的区分度展示方法，即 ROC 曲线及 AUC，对于预后模型，还可以计算模型整体区分度评价指标 C-index，我们在后面的实践中再展示其魅力吧！

1.4 净重新分类指数

NRI（Net Reclassification Index，净重新分类指数），Logistic 回归新旧模型比较只有一个 NRI 和一个 IDI，而 COX 回归针对不同时点可有多个时点的 NRI 和 IDI。

AUC 作为区分度的评价指标，虽然广泛用于预测模型的区分度评价，但其为一个综合指标，其考虑了所有预测概率作为界值的一个综合判定。而实际应用中，我们只会选取一个适宜的诊断切点，关注该切点下的诊断能力，而非所有切点构成的 AUC。

同时当我们比较两个模型的预测能力时，或者模型引入新的指标预测改善情况，我们需要一个比较两个模型预测能力的指标：NRI。

NRI 原理如下：首先将研究对象按照真实的患病情况分为两组，即患者组和非患者组，然后分别在这两个分组下，根据新、旧模型的预测分类结果（根据某个切点），整理成两个 2×2 表格，如图 1-16 所示。

患者组（N1）	新模型	
旧模型	阴性	阳性
阴性	A1	B1
阳性	C1	D1

非患者组（N2）	新模型	
旧模型	阴性	阳性
阴性	A2	B2
阳性	C2	D2

图 1-16　NRI 计算格式演示

我们主要关注被重新分类的研究对象，从图中可以看出，在患者组（总数为 N1），新模型分类正确而旧模型分类错误的有 B1 个人，新模型分类错误而旧模型分类正确的有 C1 个人，那么新模型相对于旧模型来说，正确分类提高的比例为（B1−C1）/N1，即对角线上的比例 − 对角线以下的比例。

同理，在非患者组（总数为 N2），新模型分类正确而旧模型分类错误的有 C2 个人，新模型分类错误而旧模型分类正确的有 B2 个人，那么新模型相对于旧模型正确分类提高的比例为（C2−B2）/N2，即对角线以下的比例 − 对角线以上的比例。

最后，综合患者组和非患者组的结果，新模型与旧模型相比，净重新分类指数 NRI=（B1−C1）/N1+（C2−B2）/N2。若 NRI ＞ 0，则为正改善：说明新模型的预测能力比旧模型有所改善；若 NRI ＜ 0，则为负改善，新模型的预测能力下降；若 NRI=0，则认为新模型没有改善。

然而我们上面计算的 NRI 是基于一个样本计算出来的统计量，可能是抽样误差导致的结果，因此，需要对其进行假设检验，我们可以通过计算 Z 统计量，来判断 NRI 与 0 相比

是否具有统计学显著性，统计量 Z 近似服从正态分布，公式如下：

$$Z=\frac{NRI}{\sqrt{\frac{B_1+C_1}{N_1^2}+\frac{B_2+C_2}{N_2^2}}}$$

假设某研究纳入的样本中有患者 200 例，非患者 300 例，研究者拟评价，在旧模型的基础上加入新的生物标志后，新模型预测能力的改善情况，数据如图 1-17 所示。

患者组（200）	新模型	
旧模型	阴性	阳性
阴性	20	50
阳性	10	120

非患者组（300）	新模型	
旧模型	阴性	阳性
阴性	80	40
阳性	60	120

图 1-17　NRI 演示数据

根据 NRI 计算公式：

$$NRI=（B_1-C_1）/N_1+（C_2-B_2）/N_2=（50-10）/200+（60-40）/300=26.7\%$$

代入检验公式，可得 Z=5.225，$P<0.05$，差异具有统计学意义，提示加入了新的标志物后，新模型的预测能力有所改善，正确分类的比例提高了 26.7%。

NRI 比 AUC 更加敏感，当两个模型的 AUC 差异比较无统计学显著性时，提示模型的区分能力（Discrimination）相近，但是进一步计算 NRI 后，可能会发现，新模型正确再分的能力（Reclassification）有显著提高，因此需要我们将 AUC 和 NRI 综合起来进行判断。

AUC 相当于综合实力，相当于团体赛；NRI 相当于个人的单项赛，整体比不过你，但是以某个界值为切点，我还是可以超过你的。

NRI 可以分为分类 NRI 和连续 NRI，见表 1-1，展示了分类 NRI、连续性 NRI 以及后面将介绍的 IDI 的计算。分类 NRI 需要结合专业上给的阈值，才能判定分类；连续 NRI 无须事先给出阈值，直接取新旧模型预测概率差值的符号；而 IDI 则是直接新旧模型概率的差值，同时计算时注意发生与未发生结局事件组的符号（可以简单理解为，在发生结局事件组，用新模型减去旧模型，而在未发生结局事件组，用的是旧模型减去新模型）。

表 1-1　NRI 和 IDI 模式表

编　　号	是否发生事件	原 模 型	新 模 型	阈　　值	重分类指标计算（事件组）		
					分类 NRI	连续 NRI	IDI
1	是	0.07	0.09	0.08	+1	+1	+0.02
2	是	0.09	0.07	0.08	–1	–1	–0.02
3	是	0.09	0.10	0.08	0	+1	+0.01
4	是	0.11	0.09	0.08	0	–1	–0.02.
5	否	0.07	0.09	0.08	–1	–1	–0.02

上表引自：文玲子，王俊峰，谷鸿秋 . 临床预测模型：新预测因子的预测增量值 [J]. 中国循证心血管医学杂志，2020，12(6)：655-659.

1.5 综合判别改善指数

NRI 主要用于在设定好的切点水平下判断和比较新旧模型的预测能力是否有所提高，在实际临床中容易计算，易于理解。但 NRI 的不足在于只考虑了某专业切点处的改善情况，却不能考虑模型的整体改善情况。于是一个综合判定改善情况的指标 IDI 应运而生，IDI（Integrated Discrimination Improvement，综合判别改善指数）。

IDI 的计算公式为：

IDI=(Pnew,events－Pold,events) – (Pnew,non-events－Pold,non－events)

其中，Pnew,events-Pold,events 是指患者组中，新模型预测概率的均值 – 旧模型预测概率的均值，表示预测概率提高的变化量。对于患者而言，预测概率越高，模型越准确，因此，该差值越大，提示新模型越好。

Pnew,non-events-Pold,non-events 是指非患者组，新模型预测概率的均值 – 旧模型预测概率的均值。对于非患者，预测概率越低，模型越准确，因此，差值越小则新模型越好。两部分相减，则可得 IDI。

图 1-18 中，分组代表受试对象的真实情况，PRE_1 和 PRE_2 代表旧模型和新模型的预测概率，PRE_1_1 和 PRE_2_1 代表旧模型和新模型在 0.5 切点下的分类，“0”代表未发病，“1”代表发病。利用分组分别 PRE_1_1 和 PRE_2_1 构建交叉表，就可以估算 NRI，利用分组、PRE_1_1 和 PRE_2_1，就可以计算 IDI。本章为理论，实战见后面章节。

分组	PRE_1	PRE_2	PRE_1_1	PRE_2_1
0	.14760	.04488	.00	.00
0	.09912	.05062	.00	.00
0	.65407	.32536	1.00	.00
0	.14760	.04225	.00	.00
0	.32831	.22785	.00	.00
0	.65407	.33936	1.00	.00
0	.13283	.05177	.00	.00
0	.20799	.13432	.00	.00
1	.96032	.99976	1.00	1.00
1	.89197	.98761	1.00	1.00
1	.65407	.95191	1.00	1.00
1	.34053	.12922	.00	.00
1	.49007	.99314	.00	1.00
1	.34053	.67778	.00	1.00

图 1-18　NRI 和 IDI 资料格式

IDI 也可以通过 Z 统计量来检验，以判断 IDI 与 0 相比是否具有统计学显著性，统计量 Z 近似服从正态分布，公式如下：

$$Z=\frac{IDI}{\sqrt{(SE_{\text{events}})^2+(SE_{\text{non-events}})^2}}$$

其中 SEevents 为 Pnew,events-Pold,events 的标准误，首先在患者组，计算新、旧模型对每个个体的预测概率，求得概率的差值，再计算差值的标准误。同理，SEnon-events 为 Pnew,non-events-Pold,non-events 的标准误，是在非患者组，计算新、旧模型对每个个体的预测概率，求得概率的差值，再计算差值的标准误即可。

1.6　校准度

校准度（calibration）是评价一个预测模型预测未来某个个体发生结局事件概率准确性的重要指标，反映模型预测风险与实际发生风险的一致程度，也称为一致性。可以如下方式理解区分度和校准度。

区分度：定性，你能预测出来吗？（预测出来你考试过不过）

校准度：定量，你预测出来和实际一样吗？（预测出考了多少分）

对于基于 Logistic 回归的诊断模型，校准度可以进行 H-L（Hosmer-Lemeshow）检验和绘制校准曲线（calibration plot）；对于基于 COX 回归的预后模型，校准度评价主要采用校准曲线，在校准曲线中，有一些可以评价的指标，包括截距、斜率、Brier score 等。

1.6.1　Hosmer-Lemeshow 检验

Hosmer-Lemeshow 检验是将模型预测概率分组，然后对各组的实际发生数与预测发生数进行卡方拟合优度检验，如果 $P > 0.05$，说明实际发生的频数与模型预测发生的频数比较吻合，说明模型预测得准，P 越大，说明模型表现越好。步骤如下：

（1）模型预测出概率。

（2）从小到大排序，10 分位数分组。

（3）计算每组实际观测数和模型预测数。

（4）计算卡方值，得到 P 值。

（5）P 越大，说明预测模型校准度越高，若 $P < 0.05$，则说明模型预测值与实际值存在一定的差异，校准度较差。

1.6.2　Calibration plot

校准图是对实际发生频数与模型预测发生频数相关性图，常用三种形式：①散点图，如图 1-19 所示；②柱状图，如图 1-20 所示；③线图或折线图，如图 1-21 所示。

对于 Logistic 回归的诊断模型，训练集和验证集的 H-L 和校准曲线都需要做；对于 COX 回归的预后模型，训练集和验证集的校准曲线都要做，如图 1-22 和图 1-23 所示。

图 1-23 上面两幅为训练集的校准曲线，图 1-23 下面两幅为验证集的校准曲线；图 1-24 左图为训练集 1、3、5 年校准曲线，图 1-24 右图为验证集 1、3、5 年校准曲线。

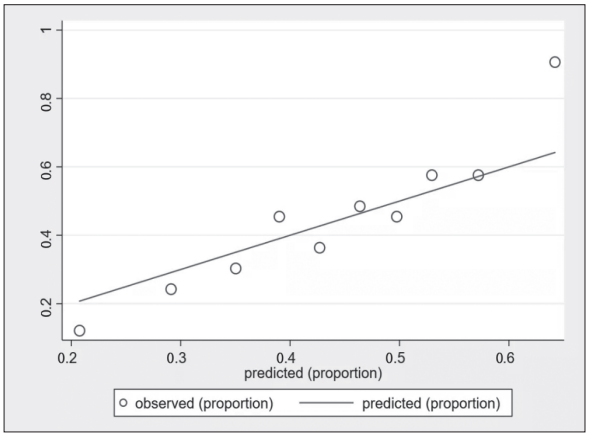

图 1-19　校准散点图

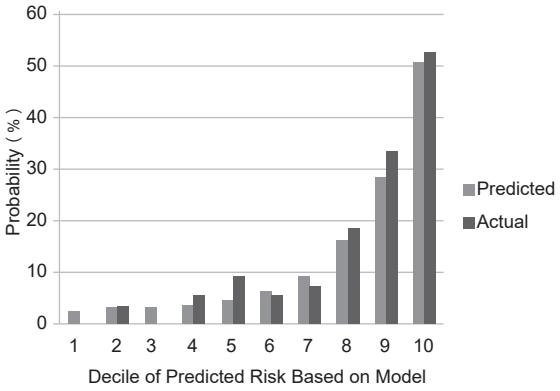

图 1-20　校准柱状图

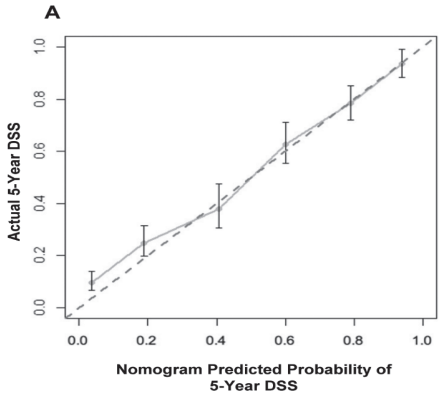

图 1-21　校准折线图

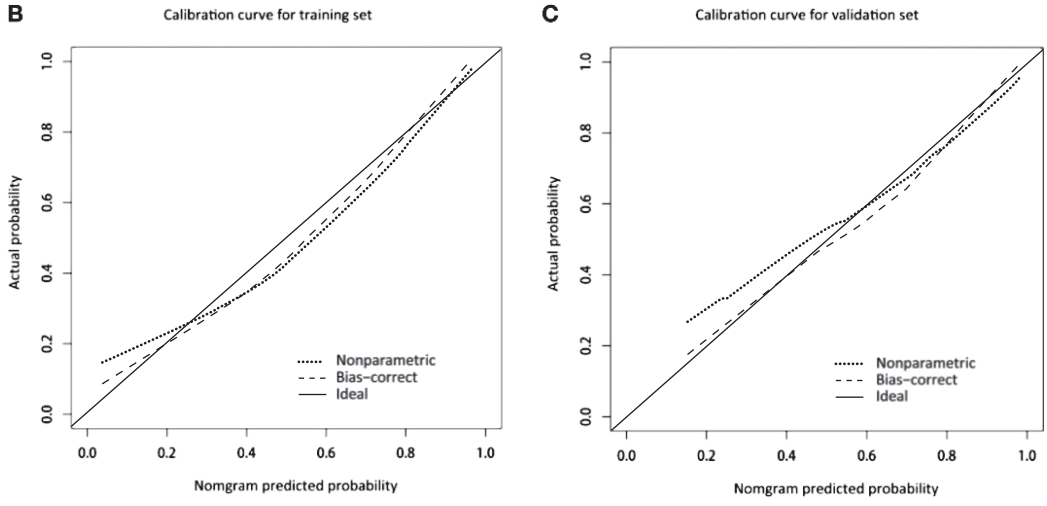

图 1-22　诊断模型的校准曲线

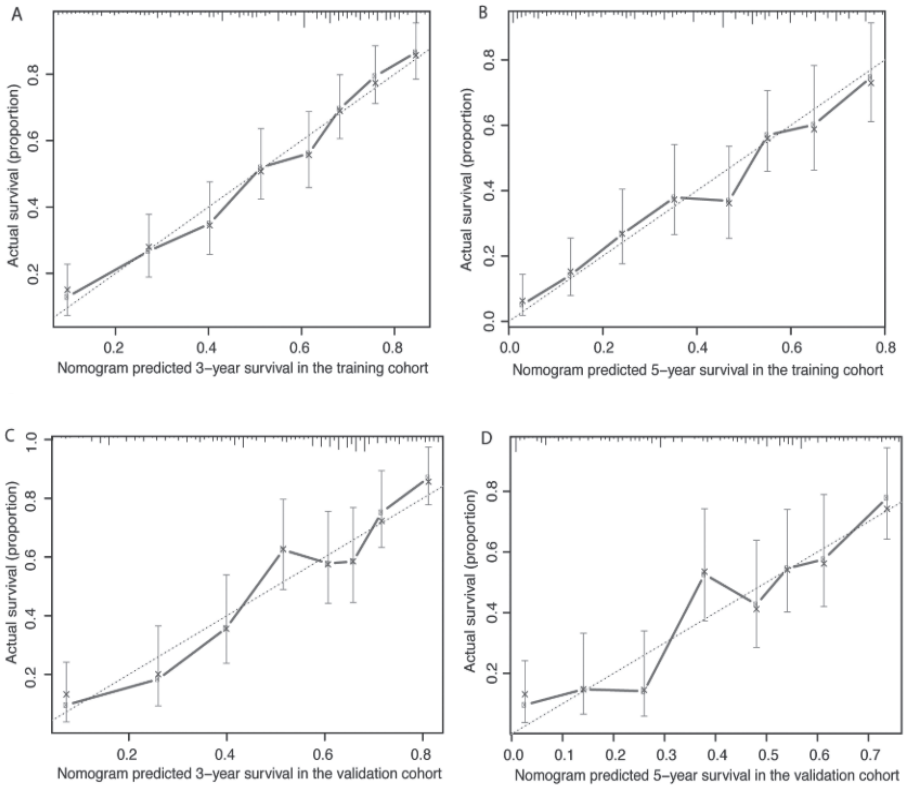

图 1-23　预后模型的校准曲线

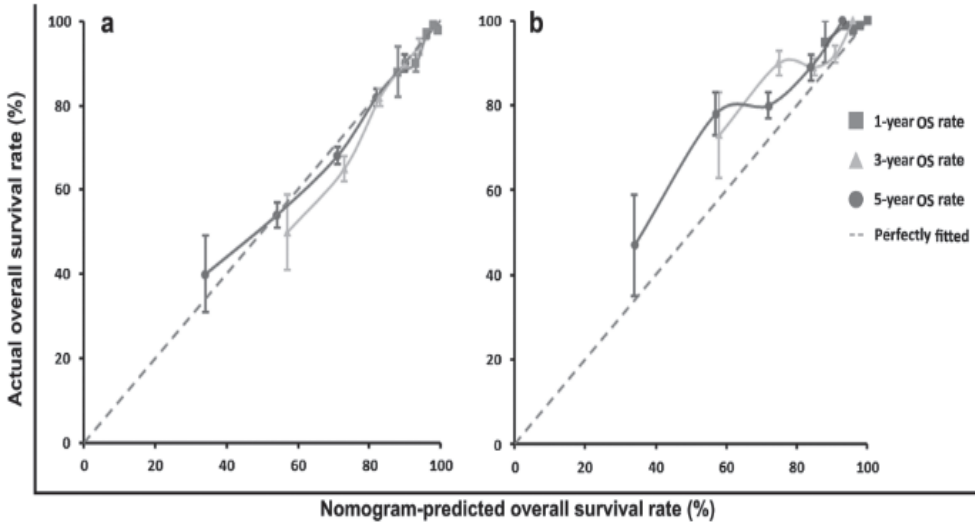

图 1-24　预后模型校准曲线（多时间点）

校准曲线上点的分布，也可以反映区分度和校准度，如图 1-25 所示。图 1-25 左上图上各点反映的是较好的区分度，但是准确度很差，因为图中两个点能够明显区分，但是不准，都没有落在参考线上；图 1-25 右上图上各点反映的是很好的准确度，但是有较差的区分度，图中各点均在参考线上，但是过于密集，区分有一定的难度。

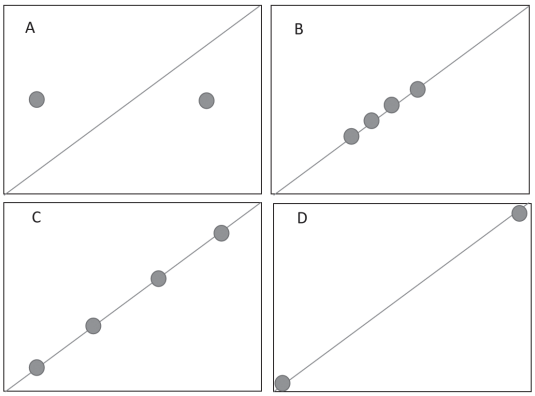

图 1-25　区分度与校准度的评价

图 1-25 左下图反映了很好的区分度和很好的准确度，点与点之间可以很好地区分，而且均在参考线上；图 1-25 右下图反映了最理想的区分和准确度，两点得到最大的区分，而且都落在参考线上。

1.7　临床决策曲线

决策曲线分析法（Decision Curve Analysis，DCA）是临床预测模型是否能够落地应用的“最后一公里”。临床上，我们通过 ROC 曲线来寻找诊断试验的切点，主要是将特异性和敏感性作为同等重要的权重进行考虑，但这种考虑真的合理吗？患者就一定受益吗？

如通过某个生物标志物预测研究对象是否患病，无论选择哪个界值，都会遇到假阳性和假阴性的可能；有时候避免假阳性受益更大，有时候则更希望避免假阴性，既然两种情况都无法避免，那就两害相权取其轻，两利相权取其重，我们就选择对患者最有利的，也就是净受益最大的方法，这就是临床效用的问题。

一个病人，如果是 X 病，做手术可延长 6 年寿命，如果不是 X 病，做手术会缩短 3 年寿命，那么某个患者经过模型预测有 40% 可能是 X 病，到底做不做手术呢？临床决策曲线就是考虑在模型各个阈值情况下，模型真正获益的净人数，如图 1-26 所示。

横坐标为阈值概率：在风险评价工具中，患者诊断为 X 病的概率为 P_i；当 P_i 达到某个阈值（P_t），就界定为阳性，采取治疗措施。此时会有病人治疗的获益，也有非病人治疗的伤害以及病人未治疗的损失。而纵坐标就是利减去弊之后的净获益（Net Benefit，NB）。

横线：代表所有样本都是阴性（$P_i < P_t$），所有人都不治疗，净获益为 0。

斜线：代表所有样本都是阳性，所有人都接受了治疗，净获益是个效率为负值的反斜线。

DCA 结果解读见图 1-27，假定选择预测概率为 40% 诊断为 X 病并进行治疗，那么每 100 人使用此模型的患者，大约有 42 人能从中获益而不损伤任何其他人的利益（其实是有人利益受损，只是 DCA 曲线分析校正了这部分人群的损失）。

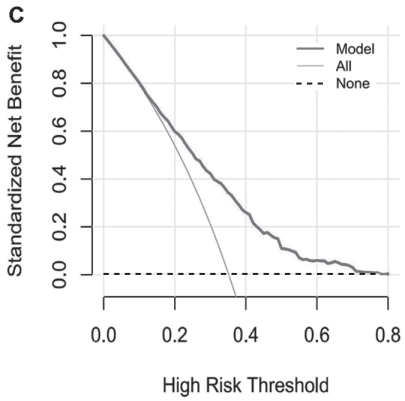

图 1-26　临床决策曲线

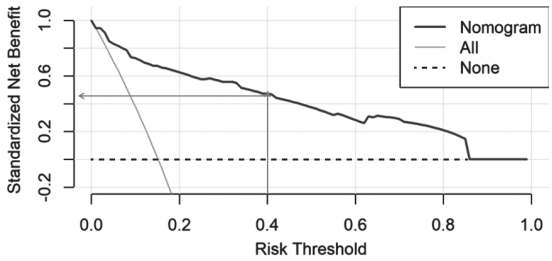

图 1-27　DCA 解读示意图

对于基于 Logistic 回归的诊断模型，训练集和验证集均需绘制 DCA 曲线，如图 1-28。而对于基于 COX 回归的预后模型，则需要绘制不同时点的 DCA 曲线，如图 1-29 所示，如 1 年、3 年、5 年等，而且也需要训练集和验证集都要绘制，如图 1-30 所示。

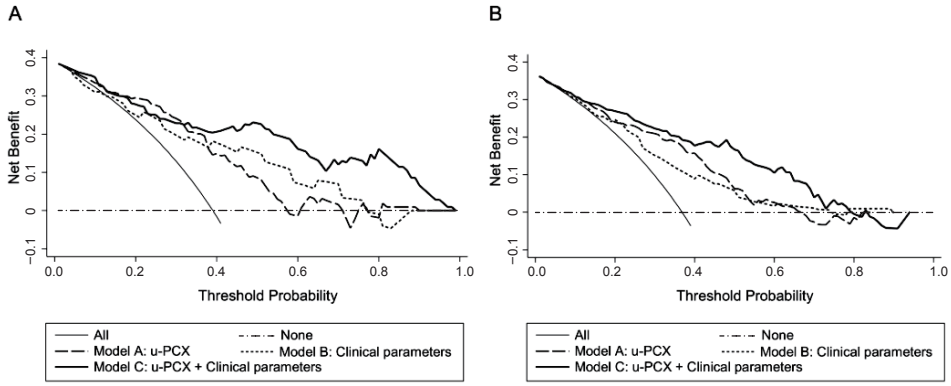

图 1-28　诊断预测模型 DCA 曲线

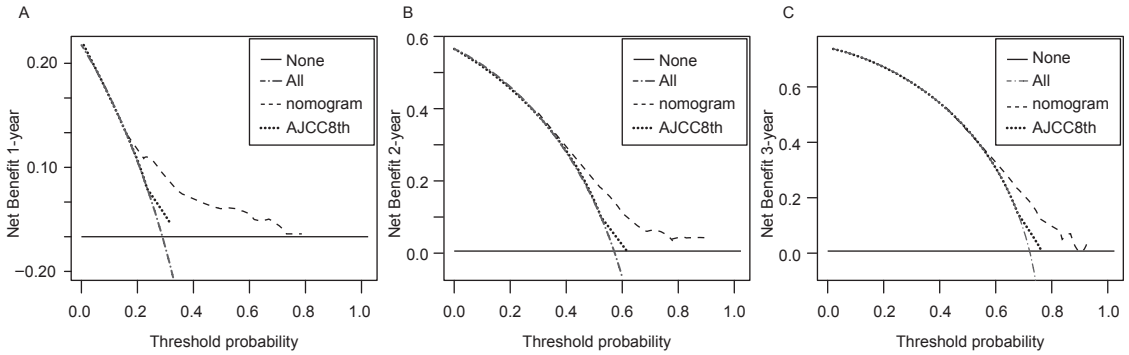

图 1-29　预后模型 DCA 曲线展示（一）

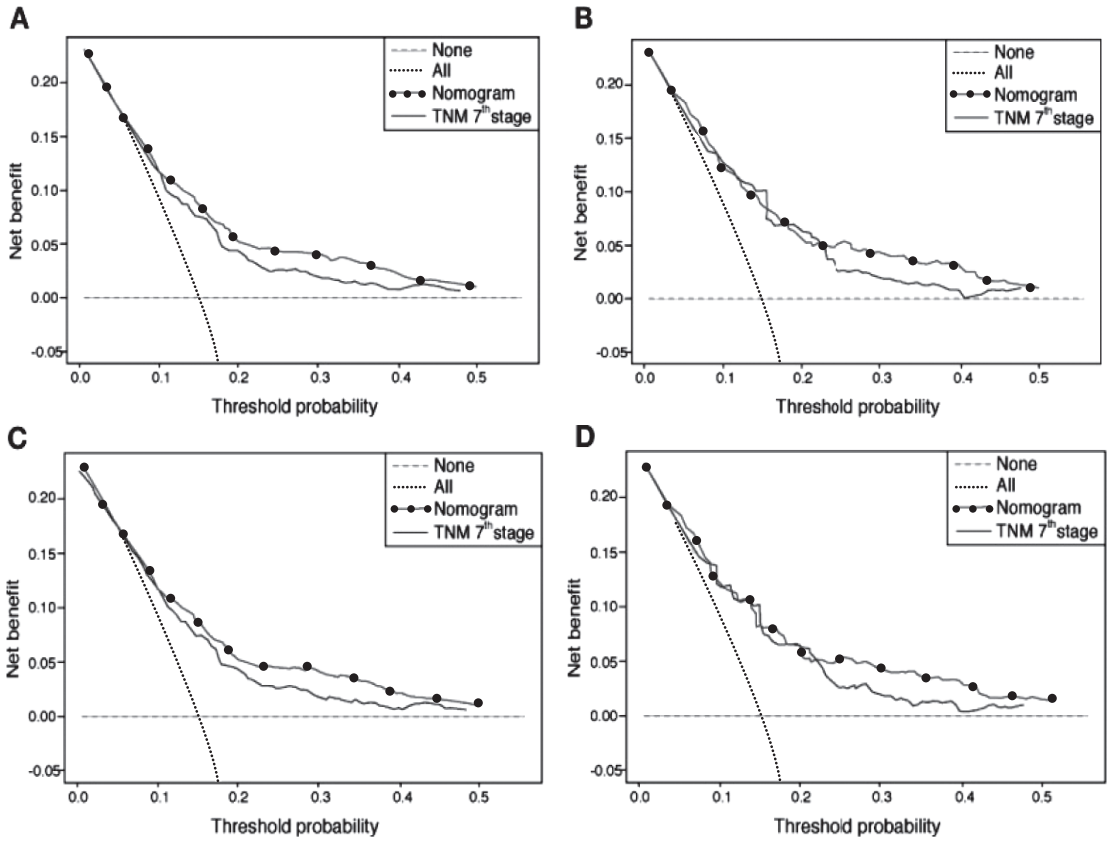

图 1-30　预后模型 DCA 曲线展示（二）

图 1-29 中 A、B、C 分别代表 1 年、2 年和 3 年的决策曲线（DCA），每张图中的两条线代表两个不同的模型。

图 1-30 中，A 和 B 代表训练集的 OS 和 CSS 的 DCA 曲线，C 和 D 代表验证集 OS 和 CSS 的 DCA 曲线。内部的两条线代表两个不同的模型。

初学临床预测模型，很多人对数据集的应用心存困惑，关于哪些指标训练集要做，哪些指标验证集要做，松哥给大家做一个总结。

训练集和验证集都要做的是：区分度（C-index 或 ROC/AUC）、校准度（校准曲线和 / 或 H-L 检验）、临床决策曲线（DCA）；Nomogram（诺莫图）只针对训练集做。而上述所有这些都是基于模型计算出来的 *P*，有 *P* 就有临床预测模型的一切。

1.8　模型可视化（Visualization）

模型可视化是指将数理模型以评分量表、网页工具或 Nomogram（诺莫图）等方式进行可视化应用。本节只介绍 Nomogram，Nomo 图非常合适在论文中展示发表。

Nomo 图的解读如图 1-31 所示，该 Nomo 图展示了一个具有 5 个因素的模型，分别为工龄、每周工作时长、年龄、文化程度和种族。

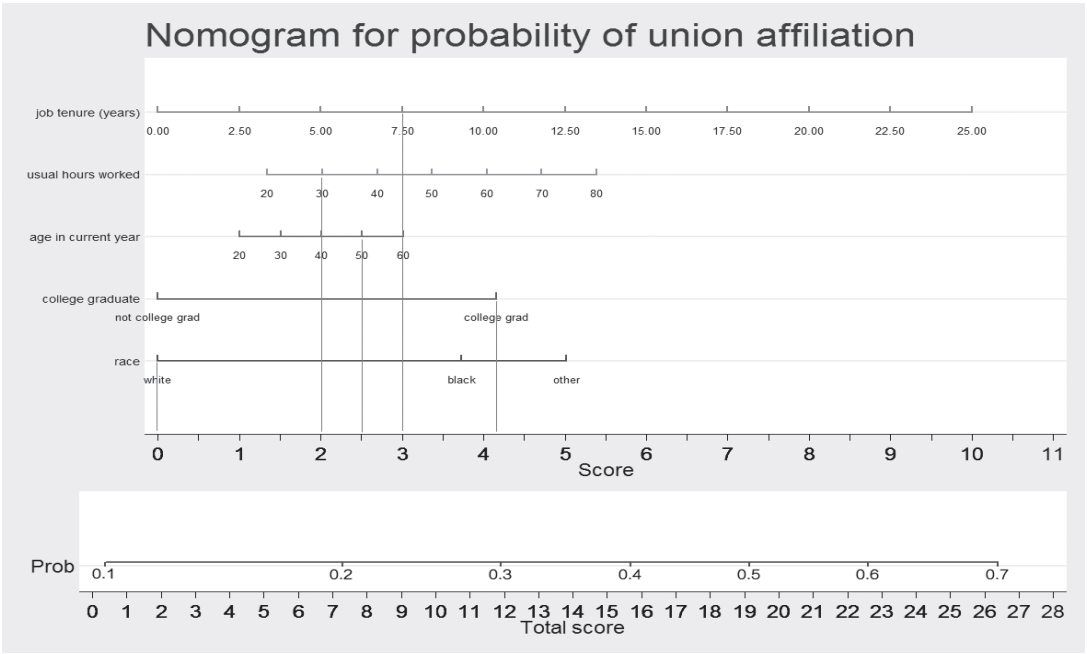

图 1-31　Nomo 图解读

如果某人工龄为 7.5 年，则对应评分为 3 分；每周工作时长为 30 小时，对应评分为 2 分；当前年龄为 50 岁，对应评分为 2.5 分；文化程度为大学毕业，对应评分为 4.2 分；种族为白色人种，对应评分为 0 分，则总共合计为 11.7 分。

图 1-32 为图 1-31 的底部部分，总分 11.7 分对应的概率约为 0.28 ～ 0.29，则说明具备上述情况的受试者，发生该病的概率为 28% ～ 29%。

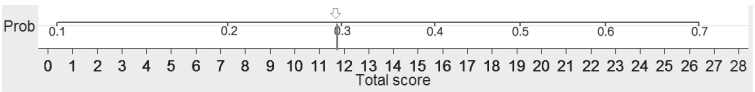

图 1-32　总分与概率对应图

Nomo 图中的评分规则，是对模型中的回归系数进行标化，具体过程稍有点复杂，好在有软件直接帮助实现，松哥也将在软件实战中讲解 Nomo 的过程。

1.9　交叉验证

交叉验证属于内部验证，往往是在我们研究的数据量不足的情况下，对建模的训练集进行分组拆分，然后进行交叉验证的过程。包括简单交叉、*K* 折交叉验证和留一法交叉验证。

1.9.1 简单交叉验证（Simple Cross Validation）

所谓的简单，是相对于其他交叉验证方法而言的。首先，我们随机地将样本数据分为两部分（比如：70% 的训练集，30% 的测试集），然后用训练集来训练模型，在测试集上验证模型及参数，如图 1-33 所示。松哥称之为“留一块”。

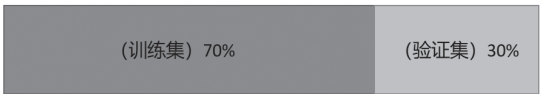

图 1-33 简单交叉验证模式图

1.9.2 *K* 折交叉验证（*K*-Folder Cross Validation）

和第一种方法不同，*K* 折交叉验证会把样本数据随机地分成 *K* 份，每次随机选择 *K*−1 份作为训练集，剩下的 1 份做测试集。当这一轮完成后，重新随机选择 *K*−1 份来训练数据。若干轮（小于 *K*）之后，选择损失函数评估最优的模型和参数，如图 1-34 所示为 10 折交叉验证模式。

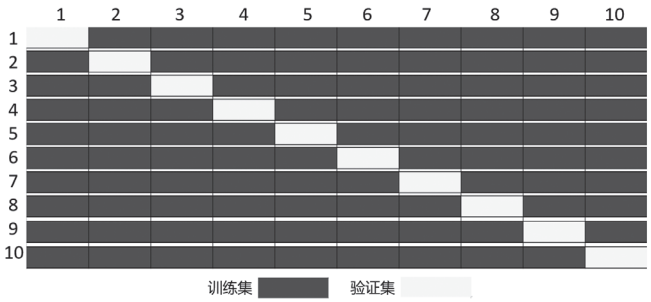

图 1-34 10 折交叉验证模式图

1.9.3 留一法交叉验证（Leave-one-out Cross Validation）

它是第二种情况的特例，此时 *K* 等于样本数 *N*，这样对于 *N* 个样本，每次选择 *N*−1 个样本来训练数据，留一个样本来验证模型预测的好坏。此方法主要用于样本量非常少的情况，比如 *N* 小于 50 时，一般采用留一法交叉验证。

一句话总结：如果我们只是对数据做一个初步的模型建立，而不做深入分析，那么简单交叉验证就可以了，否则就用 *K* 折交叉验证。在样本量少的时候，使用 *K* 折交叉验证的特例留一法交叉验证。

1.10 自助抽样法

Bootstrap 是美国斯坦福大学统计学教授 Bradley Efron 于 1979 年在论文“Bootstrap methods: another look at the jackknife”中提出。Bootstrap 的本意为靴带，Bootstrap 法来源于民间谚语：“to pull oneself up by one’s bootstrap”，说的是有一个人掉泥潭里了，然后自己拉

自己的靴带出来了，这是一个玩笑。翻译成中文就是自己帮助自己的自助法、自助抽样法、拔靴法。

简单而言，就是在建模的训练集中，进行多次有放回的小样本抽样，对每一个样本进行分析，再对若干次分析结果进行合并的分析方法，其模式图如图 1-35 所示。

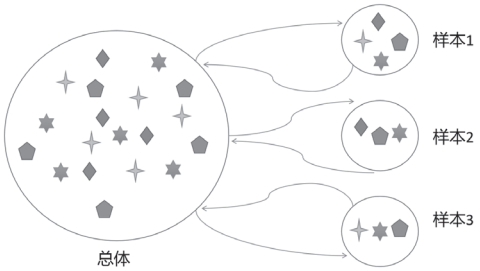

图 1-35　自助抽样法模式图

为了从单个样本中产生多个样本，Bootstrapping 作为交叉验证的一个替代方法，在原来的样本中进行有放回的随机采样，从而得到新的样本。Bootstrap 得到的样本比交叉验证的样本重叠更多，因此它们的估计依赖性更强，被认为是小数据集进行重采样的最好方法。Bootstrap 次数可以 50 次、200 次、500 次、1000 次，发表文章一般 500 次和 1000 次抽样较多。

1.11　LASSO 回归

LASSO 是由 Robert Tibshirani 于 1996 年首次提出的，全称 Least absolute shrinkage and selection operator，是一种采用了 L1 正则化（L1-regularization）的线性回归方法。该方法是一种压缩估计。它通过构造一个惩罚函数得到一个较为精练的模型，由此压缩一些回归系数，即强制系数绝对值之和小于某个固定值，同时设定一些回归系数为零，因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。下面用一个故事来讲解一下。

这是一个人类触碰“黑暗森林法则”而引发的故事：

公元 3275 年，地球科技异常发达，但是对于外星人而言，我们还是不堪一击的低等生物。由于人类对宇宙好奇，不断向外太空发射信号，期望寻找地外文明。

在宇宙中有个卡拉达星球，因为卡拉达人的资源掠夺和战争，该星球已经不再适合居住，他们便派出涉猎人在宇宙中寻找适合居住的星球。某一天，一个外星涉猎人在银河系附近，突然收到来自人类的“问候”信号，便立马将信号的宇宙坐标汇报回卡拉达星球，并请求卡拉达星球立即派出星际战舰来征服地球，然而这段通信信号，也同样被地球文明接收到。

地球立马召开多国首脑会议，商讨应对策略。最终决定，逃离地球，全部移民到人类于 3015 年发现的斯塔特星球。斯塔特星球生存环境恶劣，地球人将死囚犯送至该星球，让其自生自灭，然而大难临头之际，人类不得不逃往这一星球。

由于时间紧迫，人类的运载能力只能将 3 万人运达斯塔特星球，虽然多国总部给各国一定的指标，一些人获得消息，也暴力前往星际通航基地，最终有 10 万人进入基地。为了防止更多的人进入基地，基地开启封锁隐藏模式，与外界隔绝。

然而这 10 万人也远远超过运载能力，战舰舰长必须做出决策，让其中 3 万人乘坐战舰离开，然而这 10 万人并不知道，战舰可以带多少人离开。

在喧闹争吵之中，舰长在战舰内决定采用惩罚规则，对着 10 万人进行筛选，通过广播传达登舰规则。

第一条：凡是登舰者，必须交纳 1000 万元星际交通费，如果后面规则不符合，1000 万元不退。此规则直接筛掉 2 万人，尚有 8 万人。

第二条：有血缘关系的人，只能 1 人入选。此条颁布出，经过漫长的等待，各个有血缘关系的家族，各自推举 1 人登舰，此时还剩 6 万人。

第三条：夫妻二人者，只能一人登舰。还剩 5 万人。

第四条：愿意登舰者，达到斯塔特星球，必须做 3 年苦役！此时还剩 3 万人。

第五条：愿意登舰者，签署协议，如果航行过程中，食物短缺，愿意被随机抽中，作为大家的食物（有点残忍），还剩 1 万人。

第六条：为了表示诚信，先自愿剁去 1 个小指。还剩 5000 人。

第七条：自愿再剁去 1 个小指者，可以先进入候舰大厅，还剩 1000 人。

第八条：自愿承受一次休克式单击，能够自行苏醒者，直接登舰。此时没有人同意，最终 0 人入选。

LASSO 回归是一种带有惩罚规则的回归，不是你想加入模型就可以加入的，进入就必须付出代价。对所有候选 X 实施逐渐加大的惩罚，以让 X 表忠心，最终让每个 X 牺牲自己的生命，也就是 X 的回归系数为 0。

在逐渐加大惩罚的过程中，Y 也就测试出了每个 X 对自己的忠心程度。如图 1-36，左边喧闹烦嚣，随着惩罚加重，逐个陨落，X 的系数逐渐投向 0 的怀抱，最终在最右边，陷入一片死寂，大家终究是个 0。

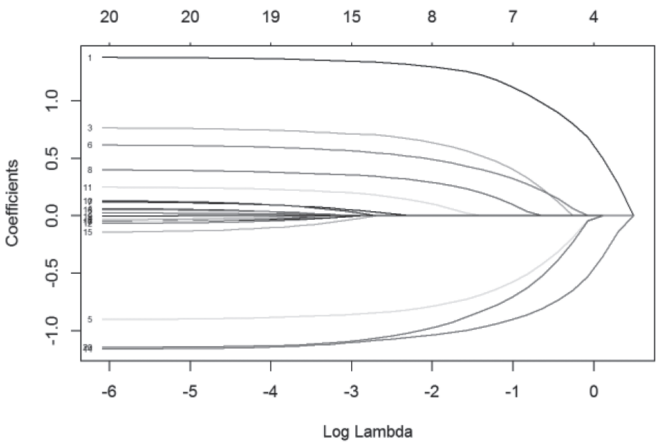

图 1-36　LASSO 路径图

然而，在此过程中，Y 就对 X 的忠心一目了然。越右边的对自己就越忠心，所以，如果要移民他星，要带着自己的团队，那么根据自己能带几个人的能力，从右边往左边选就可以了，这就是 LASSO 的变量筛选规则。

可是，选团队，带团队，并不是完全靠对自己忠心的，而是既忠心，同时这几个人组合起来的战斗力还要最强。

那么这么一来，这个团队就有若干种组合，我们会排列出几种组合，然后 PK，选取一种最优组合，而这就是 LASSO 交叉验证（CV-LASSO）。

通常我们会进行 5 重交叉，或者 10 重交叉。讲简单点就是进行 5 种或 10 种组合，然后找到最佳。

如何寻找呢？我们会做一个交叉验证的图，图中 MSE 就是误差的意思，我们可以找到误差最小的点（图 1-37 左侧虚线）。这个最小点的组合，就是我们要确定的最终入选的团队。

有人认为按照这个最小的点来选，有点没有人性，太严格了，就像考试得了 59.5 分都不给过一样，于是适当放大了 1 个 SE，如图 1-36 的右侧虚线。

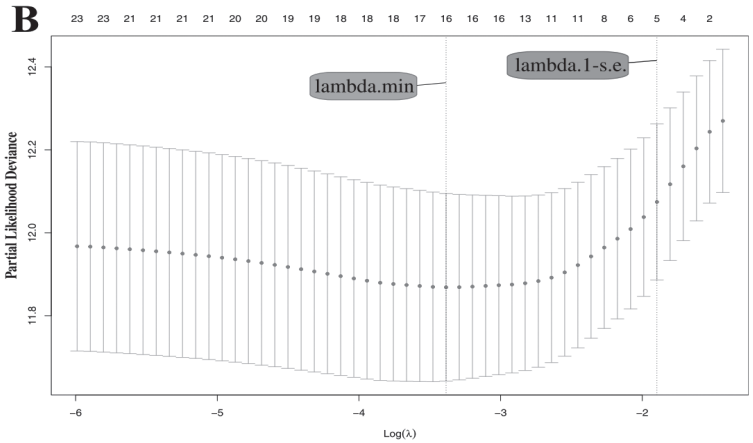

图 1-37　CV-LASSO（LASSO 交叉验证）

这就是 LASSO。统计学习应先学思想，再学软件操作，LASSO 是一种加载，就像你买了车，自行安装导航一样。所以，我们可以进行线性回归的 LASSO，逻辑回归的 LASSO，COX 回归的 LASSO，Probit 回归的 LASSO，等等。这点很重要，但很多人弄不清楚。

另外，采用 LASSO 之后，必须交叉验证，这两套技术是双胞胎，不然你无法选择出最优的团队！

1.12　临床预测模型报告规范

预测疾病状态（诊断）或未来疾病进程（预后）的发生概率对于开展个体化的精准诊疗尤为重要，临床预测模型的应用是促成临床研究转化为临床实践的有利途径。临床专家协作组于 2015 年发布了《个体预后或诊断的多变量预测模型透明报告》（*Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis*，TRIPOD），用以规范预测模型的报告过程。

TRIPOD 报告规范包括标题和摘要、前言、研究方法、结果、讨论和其他信息 6 个部分，共 22 个条目，适用于模型开发、验证、增量值等研究类别，具体清单如表 1-2 所示。

表 1-2　TRIPOD 清单

部　分	条　目	建立 / 验证	条 目 细 节
标题与摘要			
标题	1	D;V	应明确为：多因素预测模型的建立和 / 或验证，目标人群，以及预测结果
摘要	2	D;V	摘要研究对象、研究设计、研究场所、研究对象、样本量、预测指标、结局指标、统计分析方法、结果和结论
前言			
背景和目标	3a	D;V	阐述研究背景（包括诊断和预后）和多因素预测模型建立和验证的原理，并参阅已有的预测模型
	3b	D;V	明确研究目标，包括本研究是建立模型还是验证模型，还是两者都有
研究方法			
数据来源	4a	D;V	描述研究设计或数据来源（如 RCT、 队列或注册研究数据等），并应分别描述建立和 / 或验证模型的数据集
	4b	D:V	明确研究的日期信息，包括开始时间、结束时间和随访时间（如果有随访）
研究对象	5a	D;V	明确研究对象的关键信息（如初级医疗保健、二级医疗保健或普通人群等，也包括研究中心的数量和位置）
	5b	D;V	描述研究对象的入选和排除标准
	5c	D;V	如果干预与模型有关，详述干预的细节
结局指标	6a	D;V	清晰定义预测模型的预测结局，包括如何测量以及何时测量预测指标
	6b	D;V	报告对预测结局盲法评价的所有细节
预测指标	7a	D;V	清晰定义预测模型用到的所有指标，包括如何测量以及何时测量预测指标
	7b	D;V	报告对预测结局盲法评价的所有细节
样本量	8	D;V	解释研究的样本量是如何确定的
缺失数据	9	D;V	描述缺失数据的处理手段（如仅分析完整数据、单一插补和多重插补等）和插补方法的细节
统计方法	10a	D	描述统计分析中预测指标的处理手段
缺失数据	9	D;V	描述缺失数据的处理手段（如仅分析完整数据、单一插补和多重插补等）和插补方法的细节
统计方法	10a	D	描述统计分析中预测指标的处理手段
	10b	D	明确模型类型、建立过程（包括预测指标的选择）和内部验证方法
	10c	V	描述模型验证中的预测值的计算方法
	10d	D;V	明确模型预测效果的评估方法，并可比较不同的预测模型
	10e	V	描述验证模型后模型的更新（如校正等）

续表

部　分	条　目	建立 / 验证	条目细节
风险分层	11	D;V	如果做了风险分层，提供风险分层的细节
数据集比较	12	D	识别验证数据集与建立数据集在研究场所、入排标准、结局指标和预测指标上的任何不同
结果			
研究对象	13a	D;V	描述研究对象参与研究的过程，包括有结局和无结局的研究对象的数量以及随访情况，建议画流程图
	13b	D;V	描述研究对象特征，包括人口学特征、临床指标和预测指标情况以及预测指标和结局指标缺失的研究对象数量
	13c	V	比较验证数据集和建立数据集在重要变量上的分布差异，如人口学资料， 预测指标和结局指标
模型建立	14a	D	明确每次分析的研究对象和结局事件的数量
	14b	D	如果做了分析，可报告未校正的每个候选预测指标与结局指标的关系
模型规范	15a	D	提供可对个体进行预测的完整预测模型，如所有的回归系数、模型常数项或某时点的基线生存情况等
	15b	D	阐述如何使用预测模型
模型表现	16	D;V	报告预测模型的预测效果参数和可信区间
模型更新	17	V	如果做了更新，报告模型更新的结果，即模型参数和模型预测的结果
讨论			
局限性	18	D;V	讨论研究的局限性，如非随机抽样、预测指标平均事件数不足、缺失数据等
阐释	19a	V	与建立数据集或其他预测数据集的预测效果对比，讨论验证数据集的预测效果等
	19b	D;V	结合研究目的、局限性、其他类似研究结果或相关证据 , 对研究结果进行总体讨论
预示	20	D;V	讨论模型的潜在临床应用和未来研究设想
其他信息			
补充信息	21	D;V	提供补充资料和信息，如研究方案、网页计算器和数据集
资助	22	D;V	提供研究资金来源和资助方在本研究中的角色

原文可以在医学论文报告规范组织机构：EQUATOR Network（提高卫生研究质量和透明度协作网）查询，该组织以 CONSORT 工作组为框架，在全球推广使用各种医学研究报告规范，提高论文质量和透明度，促进卫生研究质量提升，在该网站几乎可以获得所有医学研究论文的报告规范。