

• 论 著 • DOI:10. 3969/j. issn. 1672-9455. 2018. 05. 019

# 系统性红斑狼疮自身抗体谱数据的解读与疾病模型预测

彭 玲<sup>1</sup>, 杨渝伟<sup>1</sup>, 王开正<sup>2△</sup>

(1. 四川省绵阳市中心医院检验科 621000; 2. 西南医科大学附属医院检验科, 四川泸州 646000)

**摘 要:**目的 通过数据挖掘建立系统性红斑狼疮(SLE)的最佳预测模型,以及自身抗体谱检测报告标准化解读数据库,以提高诊断效率。方法 收集绵阳市中心医院 2015 年 1 月至 2016 年 9 月首次送检自身抗体谱检测且诊断明确的受试者 8 904 例,其中 SLE 患者 668 例,其他自身免疫性疾病(AID)1 279 例,非 AID 6 957 例。通过受试者工作特征曲线(ROC)分析,从年龄、性别及 16 种自身抗体中筛选出对 SLE 具有预测价值的指标后,分别采用决策树、Logistic 回归和人工神经网络(ANN)建立 SLE 预测模型,并选择最优模型,建立其不同预测结果下的验后概率、误诊率等性能指标的扩展数据库。结果 经 ROC 分析,年龄、性别以及抗核抗体(ANA)、抗干燥综合征 A 抗体(SSA)、抗核糖核蛋白抗体(nRNP/Sm)、抗干燥综合征 Ro-52 抗体(Ro-52)、抗组蛋白抗体(Histone)、抗核小体抗体(Nucleosome)、抗核糖体抗 P 蛋白抗体(Rib · P)、抗双链 DNA 抗体(ds-DNA)、抗 Sm 抗体(Sm)、抗干燥综合征 B 抗体(SSB)和抗线粒体 M2 抗体(AMA-M2)均存在一定的诊断价值( $P<0.01$ )。3 种预测模型中,Logistic 模型优于其他两种模型以及任何单项自身抗体检测,差异有统计学意义( $P<0.05$ )。结论 自身抗体谱联合性别、年龄因素建立 Logistic 模型,并根据相应地区的患病率建立预测性能指标数据库,可有效提高疾病的诊断效率。

**关键词:** 系统性红斑狼疮; 自身抗体谱; 数据挖掘; 预测模型

中图法分类号: R446

文献标志码: A

文章编号: 1672-9455(2018)05-0635-04

## Interpretation and predictive modeling of antibody spectrum data for system lupus erythematosus

PENG Ling<sup>1</sup>, YANG Yuwei<sup>1</sup>, WANG Kaizheng<sup>2△</sup>

(1. Department of Clinical Laboratory, Mianyang Central Hospital, Mianyang, Sichuan 621000, China;

2. Department of Clinical Laboratory, Hospital Affiliated Southwest Medical University, Luzhou, Sichuan 646000, China)

**Abstract:** **Objective** To establish the optimum predictive model of systemic lupus erythematosus (SLE) and database for the standardized interpretation of the autoantibodies test report by data mining, so as to improve the diagnostic efficiency. **Methods** Autoantibodies test results of the 8 904 subjects who were detected at first time and had a definite diagnosis were collected from January 2015 to September 2016, including 668 cases of SLE patients, 1 279 cases of other autoimmune disease (AID) patients and 6 957 cases of non-AID patients. ROC curve analysis was used to screen valuable indexes for diagnosis of SLE from age, sex and 16 kinds of autoantibodies, followed by application of decision tree, Logistic regression and artificial neural network (ANN) to establish predictive models of SLE respectively. The optimal model was selected, and the extended database of performance indicators such as posterior probability, misdiagnosis rate and missed diagnosis rate was established for interpretation of the antibodies test report. **Results** According to the analysis of ROC curve, age, gender, ANA, SSA, nRNP/Sm, Ro-52, Histone, Nucleosome, Rib · P, ds-DNA, Sm, SSB and AMA-M2 showed application value in different degrees ( $P<0.01$ ). The Logistic model was better than the other two models, as well as any single antibody test, and the difference was statistically significant ( $P<0.05$ ). **Conclusion** The Logistic model established by combining the antibodies with the sex and age factors and the prediction performance index database established according to the local prevalence, could effectively improve the diagnostic efficiency of the disease.

**Key words:** systemic lupus erythematosus; autoantibodies; data mining; prediction model

自身抗体是自身免疫应答和自身免疫性疾病(AID)的重要特征,因此自身抗体谱检测是目前临床

广泛采用的筛查 AID 的重要手段之一。本室开展的 EUROIMMUN 自身免疫抗体谱检测共含 16 种自身

抗体,分别是抗核抗体(ANA)、抗核糖核蛋白抗体(nRNP/Sm)、抗 Sm 抗体(Sm)、抗干燥综合征 A 抗体(SSA)、抗干燥综合征 Ro-52 抗体(Ro-52)、抗干燥综合征 B 抗体(SSB)、抗拓扑异构酶 I 抗体(Scl-70)、抗 PM-Scl 抗体(PM-Scl)、抗细胞浆组酰-tRNA 合成酶抗体(Jo-1)、抗着丝点抗体 B(CENP B)、抗增殖细胞核抗原抗体(PCNA)、抗双链 DNA 抗体(ds-DNA)、抗核小体抗体(Nucleosome)、抗组蛋白抗体(Histone)、抗核糖体 P 蛋白抗体(Rib·P)和抗线粒体 M2 抗体(AMA-M2)。这些自身抗体对多种 AID,如系统性红斑狼疮(SLE)、干燥综合征、多发性肌炎、混合型结缔组织病、原发性胆汁性肝硬化等均具有良好的诊断价值<sup>[1]</sup>。但是,这些自身抗体中,除极少数如 Sm、Rib·P 为疾病标记性抗体(即仅在某一种 AID 中出现,但其敏感性低)外,其余自身抗体均在多种 AID 中呈现不同程度的检出率<sup>[2]</sup>。因此,自身抗体本身的多样性和分布的复杂性增加了自身抗体谱检测报告解读的困难,同时该报告蕴含的大信息量也进一步给报告解读带来一定难度。本文通过回顾大量样本自身抗体谱的检测数据,联合多参数建立预测模型,旨在提高 SLE 诊断效率,同时为报告的规范化解读提供依据。

## 1 资料与方法

**1.1 一般资料** 收集 2015 年 1 月至 2016 年 9 月本院行 EUROIMMUN 自身免疫抗体谱检验的所有患者 9 585 例。因诊断不明或重复送检排除 681 例,最后纳入研究 8 904 例,其中男 3 287 例,女 5 617 例,男女比例为 1:1.71。疾病分布情况:非 AID 6 957 例, SLE 668 例,类风湿关节炎 489 例,干燥综合征 141 例,系统性硬化症 79 例,肌炎/皮炎 75 例,混合型结缔组织病 58 例,未分化结缔组织病 100 例,自身免疫性肝炎 45 例,其他 AID 292 例。AID 诊断均符合美国风湿病协会或欧洲抗风湿病联盟诊断标准。根据是否 SLE 将所有患者分为两大类,然后采用分层随机抽样原则,按 4:1 比例在 SLE 患者和非 SLE 患者中分别抽样,建立训练组和验证组。

**1.2 方法** 所有受试者于清晨空腹抽取静脉血约 5 mL,室温放置 30 min 后 3 000 r/min 离心 10 min。离心后,取血清进行自身抗体检测。所有检测严格按照试剂说明书进行操作,试剂均由德国 EUROIMMUN 医学实验诊断公司生产。ANA 检测采用间接免疫荧光法,结果以阴阳性表示。每一批次试验均设立阳性和阴性对照,其余自身抗体检测采用免疫印迹法,每一检测膜条自带质量控制带,印迹膜条经 EURO 印迹法自动操作仪进行判读,结果以阴性(-)、可疑阳性(±)、+、++和+++表示。

## 1.3 SLE 预测模型的建立

**1.3.1 数据预处理** 性别分类以数字代替(男=1,女=0);自身抗体谱检测结果-、±、+、++、+++

依次转换为 0、1、2、3、4。SLE 与非 SLE 分别以“1”和“0”组表示。

**1.3.2 模型相关指标的选择** 采用 ROC 分析训练组患者中性别、年龄以及 16 种自身抗体对 SLE 的应用价值,选择有诊断意义的指标纳入模型建立。

**1.3.3 决策树模型** 对训练组数据采用 10 倍交叉验证建立最优分类树模型,增长方法选择 CHAID 法,最大树深度设为 3,父节点中的最小个案设为 100,子节点中的最小个案设为 50,生成分类规则并保存。

**1.3.4 Logistic 模型** 对训练组数据采用二元 Logistic 回归,建模方式选择向后条件法,分类标准值设为 0.5,最大迭代次数设为 20 次。通过对各变量在方程中的估计系数值进行相关性分析,从而进行各变量之间共线性诊断,对相关关系大于 0.8 的估计系数值进行校正,建立 Logistic 回归公式。

**1.3.5 BP 人工神经网络(ANN)模型** 训练模型:输入层 10 个节点,输出层 1 个节点,隐含层节点数根据经验公式计算,训练次数设定 1 000 次,学习率 0.01,采用反向传播算法,建立并储存人工神经网络模型。

**1.4 最佳模型预测指标数据库的建立** 计算上述 3 个模型分别在训练组和验证组患者的预测概率(适用 ANN 和 Logistic 模型)或结果(适用决策树模型)。在验证组患者中采用 Delong 检验比较各个模型间预测概率或曲线下面积(AUC)的大小,以 AUC 最大且与其他模型差异有统计学意义者为最优模型;在训练组患者中采用 ROC 曲线分析最佳模型的预测概率或结果,并根据分析输出的灵敏度和特异度结果,结合本地区 SLE 患病率,计算每个预测结果下的验后概率(Post P)、漏诊率和误诊率<sup>[3]</sup>,建立最优模型的预测指标数据库,用作 SLE 临床筛选和自身抗体谱报告解读的依据。

**1.5 统计学处理** 采用 SPSS19.0 和 Medcalc12.7 统计学软件进行处理分析。计数资料用百分数表示。诊断效能采用 ROC 分析;诊断性能的差异采用 Delong 检验比较两个 ROC 曲线 AUC 的大小,以  $P < 0.05$  为差异有统计学意义。

## 2 结果

**2.1 自身抗体谱诊断 SLE 的 ROC 分析** ROC 分析结果显示,性别、年龄、ANA、nRNP/Sm 等 13 个项目对 SLE 具有诊断价值( $P < 0.001$ )。其 AUC 从大到小依次为 ANA(0.889)、SSA(0.821)、nRNP/Sm(0.774)、Ro-52(0.735)、Histone(0.704)、Nucleosome(0.692)、Rib·P(0.685)、年龄(0.670)、ds-DNA(0.662)、性别(0.640)、Sm(0.638)、SSB(0.579)和 AMA-M2(0.537)。见表 1。

**2.2 决策树模型** 将上述 ROC 分析中具有诊断价值的 13 项指标作为变量输入,经分类树分析显示,输出的决策树模型深度为 3,叶节点数 17,终端节点数 11。进入决策树的变量包括共 6 种自身抗体:第 1 层

为 nRNP/Sm;第 2 层为 Nucleosome、ANA 和 Sm;第 3 层为 Rib·P 和 ds-DNA。疾病分类归属根据有序变量的值,逐渐展开成树状。

表 1 自身抗体诊断 SLE 的 AUC

| 检验结果变量     | AUC   | SE    | P     | 95%CI       |
|------------|-------|-------|-------|-------------|
| 性别         | 0.640 | 0.010 | <0.01 | 0.622~0.659 |
| 年龄         | 0.670 | 0.009 | <0.01 | 0.652~0.688 |
| ANA        | 0.889 | 0.006 | <0.01 | 0.874~0.899 |
| nRNP/Sm    | 0.774 | 0.012 | <0.01 | 0.750~0.797 |
| Sm         | 0.638 | 0.013 | <0.01 | 0.613~0.664 |
| SSA        | 0.821 | 0.011 | <0.01 | 0.800~0.842 |
| Ro-52      | 0.735 | 0.012 | <0.01 | 0.711~0.759 |
| SSB        | 0.579 | 0.013 | <0.01 | 0.554~0.604 |
| Scl-70     | 0.508 | 0.012 | 0.510 | 0.485~0.531 |
| PM-Scl     | 0.515 | 0.012 | 0.210 | 0.491~0.538 |
| Jo-1       | 0.499 | 0.012 | 0.930 | 0.476~0.522 |
| CENP B     | 0.503 | 0.012 | 0.803 | 0.480~0.526 |
| PCNA       | 0.516 | 0.012 | 0.170 | 0.493~0.539 |
| ds-DNA     | 0.662 | 0.013 | <0.01 | 0.636~0.687 |
| Nucleosome | 0.692 | 0.013 | <0.01 | 0.667~0.718 |
| Histone    | 0.704 | 0.013 | <0.01 | 0.679~0.729 |
| Rib·P      | 0.685 | 0.013 | <0.01 | 0.659~0.710 |
| AMA-M2     | 0.537 | 0.012 | 0.002 | 0.513~0.560 |

**2.3 Logistic 模型** 经 Logistic 回归分析显示,SLE 预测相关指标为性别、年龄、ANA、nRNP/Sm、Sm、SSA、ds-DNA、Nucleosome、Histone 和 Rib·P。各指标间的相关系数均小于 0.8,不存在变量的估计系数值间的共线性问题。方程各变量估计系数见表 2,建立方程如下:  $\text{Logit}(P) = 0.593 \times \text{性别} - 0.031 \times \text{年龄} + 1.085 \times \text{ANA} + 0.447 \times \text{nRNP/Sm} + 0.593 \times \text{Sm} + 0.440 \times \text{SSA} + 0.837 \times \text{ds-DNA} + 0.656 \times \text{Nucleosome} + 0.417 \times \text{Histone} + 0.577 \times \text{Rib} \cdot \text{P} - 4.876$ 。

**2.4 ANN 模型** 结合 Logistic 回归分析结果,选择性别、年龄、ANA、nRNP/Sm、Sm、SSA、ds-DNA、Nucleosome、Histone 和 Rib·P 共 10 项与 SLE 相关性较高的变量作为 ANN 输入节点,隐含层数 13,输出节点 1 个,学习率 0.01 时,训练网络误差相对最小,仿

真计算效能最高。训练结果存储为 Excel 模型文件以备应用。

**2.5 3 个模型预测效能比较** Logistic 模型、ANN 模型和决策树模型对验证组的预测概率或结果经 ROC 分析显示,其 AUC 大小依次为 0.963、0.932 和 0.862。经 Delong 检验比较,Logistic 模型对 SLE 预测价值显著优于 ANN 模型( $z=3.240, P<0.01$ )和决策树模型( $z=5.735, P<0.01$ ),而 ANN 模型显著优于决策树模型( $z=3.535, P<0.01$ )。进一步分析表明,Logistic 模型对 SLE 预测价值显著优于单项检测 AUC 最大者 ANA( $z=6.729, P<0.001$ )。因此 Logistic 模型为预测 SLE 的优选模型,盲法验证结果显示,其灵敏度为 81.2%,特异性为 97.6%,阳性预测值为 73.5%,阴性预测值为 98.4%,准确度为 96.4%。

**2.6 建立 Logistic 模型预测指标数据库** 绵阳地区 SLE 患病率为 6.97%(668/9 585)。根据该患病率及训练组中 Logistic 模型预测概率的 ROC 分析输出结果,分别计算每个阈值下的 Post P、误诊率和漏诊率,建立 Logistic 模型预测指标数据库,并列举了 P 为 0.000 8、0.005 6、0.010 1、0.066 2、0.126 1、0.178 4、0.218 4、0.918 2、0.962 6 和 0.994 3 时的模型,见表 4。

表 2 Logistic 模型分析

| 观察指标       | $\beta$ | SE    | OR   | 95%CI     | Wald   | P     |
|------------|---------|-------|------|-----------|--------|-------|
| 性别         | 0.593   | 0.217 | 1.81 | 1.18~2.77 | 7.499  | 0.006 |
| 年龄         | -0.031  | 0.005 | 0.97 | 0.96~0.98 | 37.939 | <0.01 |
| ANA        | 1.085   | 0.112 | 2.96 | 2.38~3.69 | 93.889 | <0.01 |
| nRNP/Sm    | 0.447   | 0.063 | 1.56 | 1.38~1.77 | 49.997 | <0.01 |
| Sm         | 0.593   | 0.149 | 1.81 | 1.35~2.42 | 15.763 | <0.01 |
| SSA        | 0.440   | 0.047 | 1.55 | 1.42~1.70 | 86.165 | <0.01 |
| ds-DNA     | 0.837   | 0.116 | 2.31 | 1.84~2.90 | 51.806 | <0.01 |
| Nucleosome | 0.656   | 0.122 | 1.93 | 1.52~2.45 | 28.701 | <0.01 |
| Histone    | 0.417   | 0.087 | 1.52 | 1.28~1.80 | 23.073 | <0.01 |
| Rib·P      | 0.577   | 0.072 | 1.78 | 1.55~2.05 | 64.934 | <0.01 |

表 3 Logistic 模型预测指标数据库

| Logistic-P | 灵敏度     | 特异度     | 阳性预测值     | Post P(%) | 误诊率(%) | 漏诊率(%) |
|------------|---------|---------|-----------|-----------|--------|--------|
| 0.000 8    | 1.000 0 | 0.000 6 | 1.000 0   | 6.974 0   | 99.94  | 0.00   |
| 0.005 6    | 0.977 4 | 0.429 9 | 1.710 0   | 11.382 8  | 57.01  | 2.26   |
| 0.010 1    | 0.969 9 | 0.672 1 | 2.960 0   | 18.141 0  | 32.79  | 3.01   |
| 0.066 2    | 0.932 3 | 0.922 3 | 12.000 0  | 47.339 8  | 7.77   | 6.77   |
| 0.126 1    | 0.857 1 | 0.948 4 | 16.610 0  | 55.446 5  | 5.16   | 14.29  |
| 0.178 4    | 0.834 6 | 0.966 6 | 24.990 0  | 65.182 9  | 3.34   | 16.54  |
| 0.218 4    | 0.827 1 | 0.970 9 | 28.380 0  | 70.957 6  | 2.49   | 18.80  |
| 0.918 2    | 0.421 1 | 0.995 1 | 86.680 0  | 86.556 8  | 0.49   | 57.89  |
| 0.962 6    | 0.360 9 | 0.998 2 | 198.140 0 | 93.758 5  | 0.18   | 63.91  |
| 0.994 3    | 0.240 6 | 0.999 4 | 396.270 0 | 96.778 7  | 0.06   | 75.94  |

**2.7 建立 SLE 的 Logistic 模型预测系统** 利用 Excel 表格的计算和查找匹配功能,将 Logistic 模型方程与预测指标数据库相联系,建立 SLE 的 Logistic 模型预测系统。只要将检验结果数字化处理后输入或导入,该系统可自动计算预测概率,然后根据预测概率自动从预测指标数据库中检索并输出该患者被诊断为 SLE 的验后概率和误诊率、漏诊率,由此提高工作效率,增加预测模型的实用性。

### 3 讨论

SLE 是 AID 中最具代表性的疾病,其发病机制复杂,临床表现复杂多样,早期发现、诊断和治疗是降低 SLE 致残或致死率最有效的方法。目前临床上较多采用 1997 年美国风湿病协会制定的 SLE 诊断标准,但其灵敏度有限,不利于早期诊断。2012 年, SLE 国际临床协作组 (SLICC) 提出的新标准,虽提高了灵敏度,但特异性有所降低,易造成误诊<sup>[1]</sup>。既往报道多有采用不同自身抗体联合检测提高 SLE 检出率<sup>[4-5]</sup>,但少有依据自身抗体建立的筛选模型。有较多研究分别运用蛋白指纹图谱技术和代谢组学方法建立 SLE 诊断模型<sup>[6-8]</sup>,为 SLE 诊断和监测开辟了新的途径,但这两种方法技术要求高,尚不能在临床常规开展。

EUROIMMUN 自身免疫抗体谱检测目前在临床实验室应用广泛,其中多种自身抗体是诊断 SLE 或其他 AID 的重要参考指标。本研究针对大样本量检测结果进行数据挖掘,从实验室角度建立 SLE 预测模型,不仅可提高 SLE 诊断效率还有助于无症状患者的早期发现并提前进行免疫干预和制订个体化治疗方案,同时联合本地区患病率建立预测指标数据库为自身抗体谱标准化解读提供客观依据。

本文中 ROC 分析显示性别、年龄以及 11 种自身抗体各单一项目对 SLE 仅有中等或较低诊断价值,而联合性别、年龄和多种自身抗体建立的 3 种数学模型,其 AUC 明显增高,Logistic 模型和 ANN 模型的 AUC 分别为 0.963 和 0.932,均显示较高的诊断效能。Logistic 模型的回归公式直观且易于理解,计算简单方便,变量纳入合理,其诊断 SLE 的灵敏度为 81.2%,特异度 97.6%,准确度 96.4%,为本研究中的优选模型。ANN 模型较高的诊断效率与既往研究结果一致<sup>[9-11]</sup>,表明其应用于疾病的诊断有较高的实用价值,但其网络训练的最佳原则、隐含层层数及节点的选取缺乏明确的指导方法,并且计算过程不易理解,在本研究中列为次选模型。决策树模型在本研究中对疾病的诊断效率相对偏低,该模型易于理解和实现,可判断各变量的相对重要性,并能从中提取诊断规则,缺点在于对具有连续值的属性预测比较困难,数据集类别太多时,误差会增加<sup>[12]</sup>。

预测指标数据库中的 Post P、误诊率和漏诊率这

3 项性能指标相较于直接输出的模型概率而言,考虑了试验灵敏度、特异度及患病率(即验前概率)等因素的影响,用于辅助临床诊断和解释检验报告比较客观、真实,为实验室开展检验信息咨询找到一个新途径。但是该数据库不具有通用性,因为验后概率的计算受验前概率的影响,而验前概率在用于预测时需要患者的来源符合本研究所推断的群体<sup>[13]</sup>,因此本研究中的预测指标数据库只适用于绵阳市中心医院的自身抗体谱报告解释,其他地区可根据本地 SLE 患病率及相关检测数据建立相应的数据库。

### 参考文献

- [1] 仲人前,杨再兴.几种常见自身免疫性疾病诊断标准进展及展望[J].实用检验医师杂志,2014,6(2):65-69.
- [2] 李永哲.自身抗体检测技术临床推广应用和质量保证工作中应重视的问题[J].中华检验医学杂志,2006,29(9):769-773.
- [3] 王开正,陶华林,王治国.检验医学信息学[M].成都:四川科学技术出版社,2004:152-168.
- [4] 姜清明.ANA、抗 ds-DNA 抗体及抗 ENA 抗体在自身免疫性疾病诊断中的应用[J].国际检验医学杂志,2016,37(14):2011-2013.
- [5] 彭赛蛟,陶锡东,王敏.不同自身抗体联合检测在系统性红斑狼疮中的诊断价值[J].中国免疫学杂志,2014,30(6):831-833.
- [6] 张妍,刘晓敏,李丽,等.系统性红斑狼疮患者血清蛋白质组学研究[J].标记免疫分析与临床,2014,21(2):184-187.
- [7] 黄卓春,石运莹,蔡培,等.运用蛋白质指纹图谱技术建立系统性红斑狼疮诊断模式的研究[J].四川大学学报(医学版),2009,40(3):499-503.
- [8] 彭赛蛟,陶锡东,王敏.不同自身抗体联合检测在系统性红斑狼疮中的诊断价值[J].中国免疫学杂志,2014,30(6):831-833.
- [9] OMID P, SARA D, HEDIEH Z, et al. A diagnostic model for cirrhosis in patients with non-alcoholic fatty liver disease: an artificial neural network approach[J]. Med J Islamic Repu, 2014, 28: 116-121.
- [10] BARIS S, MEHMET K, SONMEZ O, et al. Artificial neural network analysis for evaluating cancer risk in multinodular goiter[J]. J Res Med Sci, 2013, 18(7): 554-557.
- [11] 江永平,张勇,蒋宁,等.肿瘤标志物联合人工神经网络对大肠癌的预警[J].中国卫生检验杂志,2015,25(3):371-373.
- [12] 牟冬梅,冯超,王萍.数据挖掘方法在医学领域的应用及 SWOT 分析[J].医学信息学杂志,2015,36(1):53-57.
- [13] 杨尚瑜,杨梅,王开正,等.血糖三项检验的数据挖掘及应用[J].山东医药,2014,54(13):16-18.