

# 临床诊断决策树方法用于结核性胸膜炎 诊断的初探

孔忠顺 刘京铭 高孟秋 黄麦玲

**【摘要】 目的** 利用临床诊断决策树的方法建立结核性胸膜炎(tuberculous pleurisy, TBP)综合诊断策略,并探讨其对 TBP 的诊断价值。**方法** 采用回顾性研究方法,根据自行制定的结核性胸膜炎及恶性胸腔积液的入组标准,收集首都医科大学附属北京胸科医院 2014 年 1 月至 2015 年 12 月符合入组标准的住院患者病历资料,共 314 例,分为 TBP 组(205 例)和恶性胸腔积液组(109 例)。采用随机数的生成方法将综合数据按照 3:1 比例,分为训练样本数据集及验证样本数据集,然后进行决策树算法(CART)分析,生成结核性胸膜炎的临床诊断决策树,最后将生成的决策树模型对验证样本数据集进行验证并计算出检测效应值,以完成对 TBP 综合诊断策略的验证。**结果** 对 25 项用于构建临床诊断决策树的指标进行了单因素统计分析显示,其中有 16 项指标在 TBP 组和恶性胸腔积液组差异有统计学意义。以独立构建的临床诊断决策树工作流程为基础,进行 1000 次模拟实验,全部循环构建的 1000 棵决策树平均利用的评判因素为 $(8.57 \pm 1.63)$ 个。对实验结果进行相关评价指标的计算,结果显示临床诊断决策树用于 TBP 诊断的敏感度为 98.14%,特异度为 93.64%,符合率为 95.01%。对决策树中各项指标的贡献得分排序显示,排名前 9 项指标依次是胸腔积液腺苷脱氨酶、红细胞沉降率、发热、胸腔积液 C 反应蛋白、年龄、血结核抗体、血 T 细胞斑点试验 B、性别、乏力等。**结论** 临床诊断决策树方法是 TBP 与恶性胸腔积液有效鉴别诊断策略之一。

**【关键词】** 结核;胸膜; 诊断; 决策树

**A preliminary study of decision tree in clinical diagnosis of tuberculous pleurisy** KONG Zhong-shun, LIU Jing-ming, GAO Meng-qiu, HUANG Mai-ling. The Second Ward of TB Department, Beijing Chest Hospital, Capital Medical University, Beijing 101149, China

Corresponding author: HUANG Mai-ling, Email: huangmailing@163.com

**【Abstract】 Objective** To form comprehensive diagnosis strategy of tuberculous pleurisy (TBP) using the decision tree in the clinical, and to evaluate the value of decision tree in diagnosis of TBP. **Methods** Based on inclusion criteria of TBP and malignant pleural effusion, 314 patients from Beijing Chest Hospital affiliated to Capital Medical University between January 2014 to December 2015 were retrospectively studied. These patients were divided into TBP group (205 cases) and malignant pleural effusion group (109 cases). And the comprehensive data were randomly divided into the training sample data set and validation sample data set according with the ratio of 3:1 using random number statistics, and then a diagnosis tree for clinical diagnosis of TBP were builded, which was used to verify validation sample data and calculate detection effect value, in order to verify the comprehensive diagnosis strategy. **Results** A single factor statistical analysis was made on 25 indexes of constructing decision tree in the clinical and it was found that, 16 indexes were statistically significant between TBP group and malignant pleural effusion group. A total of 1000 simulated experiments were carried out based on the clinical decision tree, the average utilization of the 1000 decision trees constructed by the whole cycle is  $8.57 \pm 1.63$ . According to the algorithm of relevant evaluation indexes of the experimental results, the sensitivity for diagnosis of TBP by the constructed clinical decision tree was 98.14% and the specificity was 93.64%, the accuracy was 95.01%. According to the contribution of the indexes in the decision tree, the front 9 indexes were the pleural effusion, erythrocyte sedimentation rate, fever, pleural effusion C-reactive protein, age, blood tuberculous antibody, blood T lymphocyte spot test B, sex and fatigue, etc. **Conclusion** The clinical decision tree is one of the effective methods for differentiating TBP from malignant pleural effusion.

**【Key words】** Tuberculosis, pleural; Diagnosis; Decision trees

结核性胸膜炎(tuberculous pleurisy, TBP)是临床上常见的肺外结核之一,同时在某些国家也是造成胸腔积液的首要病因<sup>[1]</sup>。在结核病高流行地区,约有 50% 患者胸腔积液由 TBP 引起<sup>[2]</sup>。然而, TBP 的发病原因并非都是由结核分枝杆菌(Mtb)直接感染造成的局部炎症,大多是由 Mtb 引起的迟发变态反应所导致,很难直接在 TBP 患者的胸腔积液中直接找到 Mtb 的细菌学证据<sup>[3]</sup>。因此, TBP 的诊断目前仍是亟待解决的临床问题。笔者收集首都医科大学附属北京胸科医院 TBP 及恶性胸腔积液病例资料,采用回顾性研究的方式,利用数据挖掘和机器学习的方法建立 TBP 与恶性胸腔积液临床诊断决策树,探讨决策树方法用于 TBP 诊断的可行性。

### 资料和方法

1. 患者选择:采用回顾性研究方法,根据自行制定的结核性胸膜炎及恶性胸腔积液的人组标准,收集首都医科大学附属北京胸科医院 2014 年 1 月至 2015 年 12 月符合入组标准的住院患者病历资料,共 314 例,分为 TBP 组(205 例)和恶性胸腔积液组(109 例)。

2. 纳入及排除标准:参考文献<sup>[4]</sup>及有关文献进行制定。(1) TBP 组纳入标准:临床诊断为 TBP,同时胸腔积液有直接细菌学证据(涂片抗酸染色或培养)或胸膜活检病理有典型的特征性结核病理表现者;或临床诊断为 TBP,同时有痰细菌学证据(涂片

抗酸染色或培养),或肺部有明确结核病灶,且抗结核并抽胸腔积液治疗后胸腔积液得以吸收者。(2) 恶性胸腔积液组纳入标准:胸腔积液患者,于胸腔积液细胞学或胸膜组织病理学找到恶性细胞者。(3) 排除标准(两组患者均需排除):① 结核病同时并发恶性肿瘤;② 诊断为 TBP 或恶性胸腔积液同时并发严重肝病、肾病、心功能不全、严重低蛋白血症、结缔组织疾病、自身免疫性疾病、HIV 感染或艾滋病、长期应用免疫抑制剂等。

3. 数据处理:在数据录入之前制定收集数据项目及数据类型标准。根据数据分析预实验从患者一般情况、症状及实验室检查等 3 方面共 25 项指标纳入后期诊断模型的构建。数据录入采取独立双录入的原则,保证数据的可靠性。数据分析之前进行异常值的筛选,缺失项的复合等数据清理工作。对清洗后的数据根据分组标准分为 TBP 组及恶性胸腔积液组,并将数据总和作为综合样本数据。之后,编写相关的计算机程序,并构建分析程序工作流程。应用专门的生物信息分析平台 BioKepler<sup>[5]</sup> 及 R3.1<sup>[6]</sup> 构建临床诊断决策树分析工作流程(work flow),见图 1。该流程首先进行数据读取,之后进行数据清理,并采用随机数的生成方法将综合数据按照 3:1 比例,随机分为训练样本数据集及验证样本数据集,然后进行决策树算法(CART),生成 TBP 的临床诊断决策树,最后将生成的决策树模型对验证样本数据集进行验证并计算出检测效应值,该程序将上述流程进行了模块化,并且根据研究 workflow

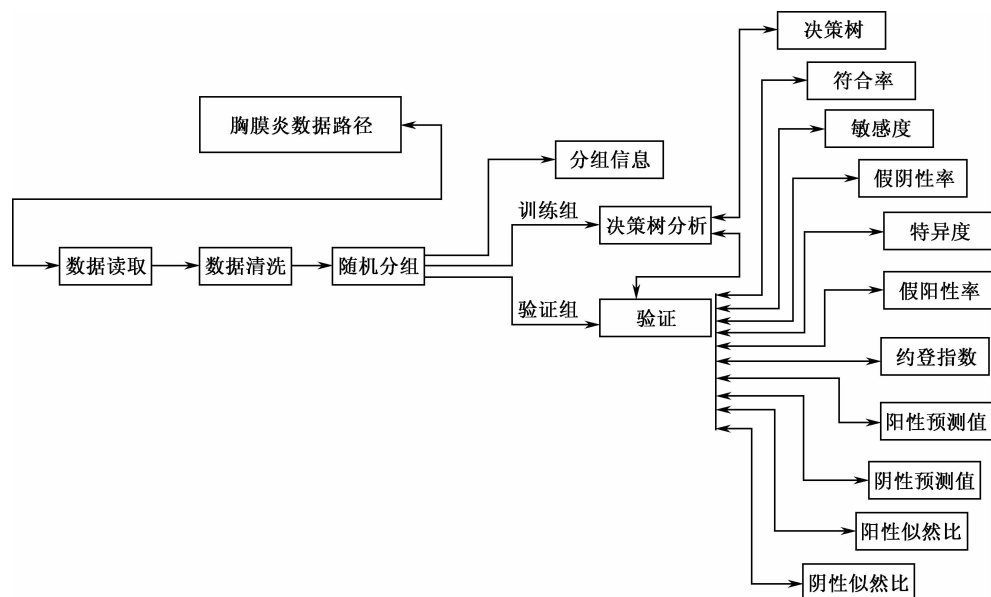


图 1 临床诊断决策树分析工作流程

程进行串联,完成标准化的目的,同时可以实际运行并直接输出相应的结果。

采用 R3.1<sup>[6]</sup> 软件完成全部决策树及统计分析。对清洗后的相关数据进行两组间各影响因素差异的统计学分析。计量资料采用  $M(Q_1 \sim Q_3)$  表示,应用秩和检验进行比较;计数资料的比较采用  $\chi^2$  检验,不满足  $\chi^2$  检验条件者采用 Fisher 精确概率法进行检验;以  $P < 0.05$  为差异有统计学意义。

## 结 果

### 1. 基本情况: TBP 组 205 例,包括男 164 例

(80.0%),女 41 例 (20.0%),平均年龄  $(46.73 \pm 19.47)$  岁。恶性胸腔积液组 109 例,包括男 65 例 (59.6%),女 44 例 (40.4%),平均年龄  $(60.92 \pm 11.10)$  岁。两组比较显示,TBP 组男性患者构成比明显高于恶性胸腔积液组 ( $\chi^2 = 14.95, P < 0.01$ ); TBP 组的平均年龄明显低于恶性胸腔积液组 ( $t = 7.02, P < 0.01$ )。在临床症状方面,共纳入 7 项指标,其中 5 项指标 TBP 组与恶性胸腔积液组比较存在差异,见表 1。

2. TBP 临床诊断决策树的生成与检验:因本次纳入研究的患者信息相对较少,会出现决策树生成

表 1 研究对象临床特征在结核性胸膜炎组和恶性胸腔积液组间的比较

临床特征	结核性胸膜炎组	恶性胸腔积液组	统计学检验值	P 值
发热			$\chi^2 = 136.85$	$< 0.01$
是	158(77.07)	8(7.62)		
否	47(22.93)	101(92.38)		
乏力			$\chi^2 = 32.17$	$< 0.01$
是	84(40.89)	11(10.09)		
否	121(59.11)	98(89.91)		
盗汗			$\chi^2 = 40.34$	$< 0.01$
是	68(33.17)	2(1.83)		
否	137(66.83)	107(98.17)		
胸闷			$\chi^2 = 9.50$	0.002
是	83(40.49)	64(58.72)		
否	122(59.51)	45(41.28)		
气短			$\chi^2 = 8.45$	0.004
是	69(33.66)	55(50.46)		
否	136(66.34)	54(49.54)		
咳嗽			$\chi^2 = 0.063$	0.802
是	91(44.39)	50(45.98)		
否	114(55.61)	59(54.02)		
胸腔积液颜色			$\chi^2 = 15.29$	$< 0.01$
黄色	188(91.71)	84(77.06)		
红色	17(8.29)	25(22.94)		
血清结核抗体			$\chi^2 = 55.04$	$< 0.01$
阳性	94(45.85)	5(4.59)		
阴性	111(54.15)	104(95.41)		
胸腔积液结核抗体			$\chi^2 = 1.48$	0.223
阳性	52(25.36)	21(19.27)		
阴性	153(74.64)	88(80.73)		

续表 1

临床特征	结核性胸膜炎组	恶性胸腔积液组	统计检验值	P 值
体质量下降值(kg)	0.00(0.00~4.00)	0.00(0.00~3.50)	$Z=-0.69$	0.491
血 T 淋巴细胞斑点试验 A(SFC/ $10^6$ )	65.00(28.00~183.00)	18.00(0.00~52.00)	$Z=-6.24$	$<0.01$
血 T 淋巴细胞斑点试验 B(SFC/ $10^6$ )	66.00(21.00~176.00)	15.00(4.00~40.00)	$Z=-6.17$	$<0.01$
血红细胞沉降率(mm/1 h)	35.40(15.20~55.50)	17.60(3.50~33.50)	$Z=-3.45$	$<0.01$
胸腔积液粒细胞数( $\times 10^3/\mu\text{l}$ )	2.16(0.95~4.95)	1.87(1.06~3.37)	$Z=-1.75$	0.081
淋巴细胞百分比(%)	80.00(56.20~87.85)	76.50(55.40~85.20)	$Z=-0.60$	0.526
胸腔积液葡萄糖(mmol/L)	5.00(3.80~6.50)	5.70(4.50~7.30)	$Z=-2.88$	$<0.01$
胸腔积液蛋白(g/L)	48.50(43.50~52.40)	45.50(40.90~52.60)	$Z=-0.09$	0.916
胸腔积液氯化物(mmol/L)	105.40(102.40~107.40)	106.70(103.70~109.80)	$Z=-1.77$	0.087
胸腔积液腺苷脱氨酶(U/L)	22.83(8.58~39.98)	7.35(2.49~17.92)	$Z=-10.43$	$<0.01$
胸腔积液 C 反应蛋白(mg/L)	44.50(23.25~56.05)	12.10(8.60~15.40)	$Z=-5.35$	$<0.01$
胸腔积液乳酸脱氢酶(U/L)	323.00(197.00~589.00)	314.90(164.00~512.00)	$Z=-1.77$	0.842
胸腔积液 T 淋巴细胞斑点试验 A(SFC/ $10^6$ )	128.00(6.00~366.00)	116.50(0.00~312.00)	$Z=-0.07$	0.921
胸腔积液 T 淋巴细胞斑点试验 B(SFC/ $10^6$ )	122.00(4.00~286.00)	116.00(0.00~298.00)	$Z=-0.37$	0.737

注 表中计数资料括号外数值为“患者例数”,括号内数值为“构成比(%)”;计量资料括号外数值为中位数(M),括号内数值为第 25 百分位数~第 75 百分位数( $Q_1 \sim Q_3$ )

不稳定的情况。为客观的评价决策树策略用于 TBP 诊断的可行性,对已构建的临床诊断决策树分析工作流程进行 1000 次循环模拟,全部循环构建的 1000 棵决策树平均利用的评判因素为  $(8.57 \pm 1.63)$  个。以约登指数(Youden index)进行从大到小排列,将约登指数为 1 的前 5 棵决策树以图的形式显示于图 2。

3. 决策树的检验:应用经 1000 次循环构建的决策树对验证组进行预测,对预测结果进行了 10 项临床试验检测指标的验证,10 项临床试验检测指标的分布情况见表 2。

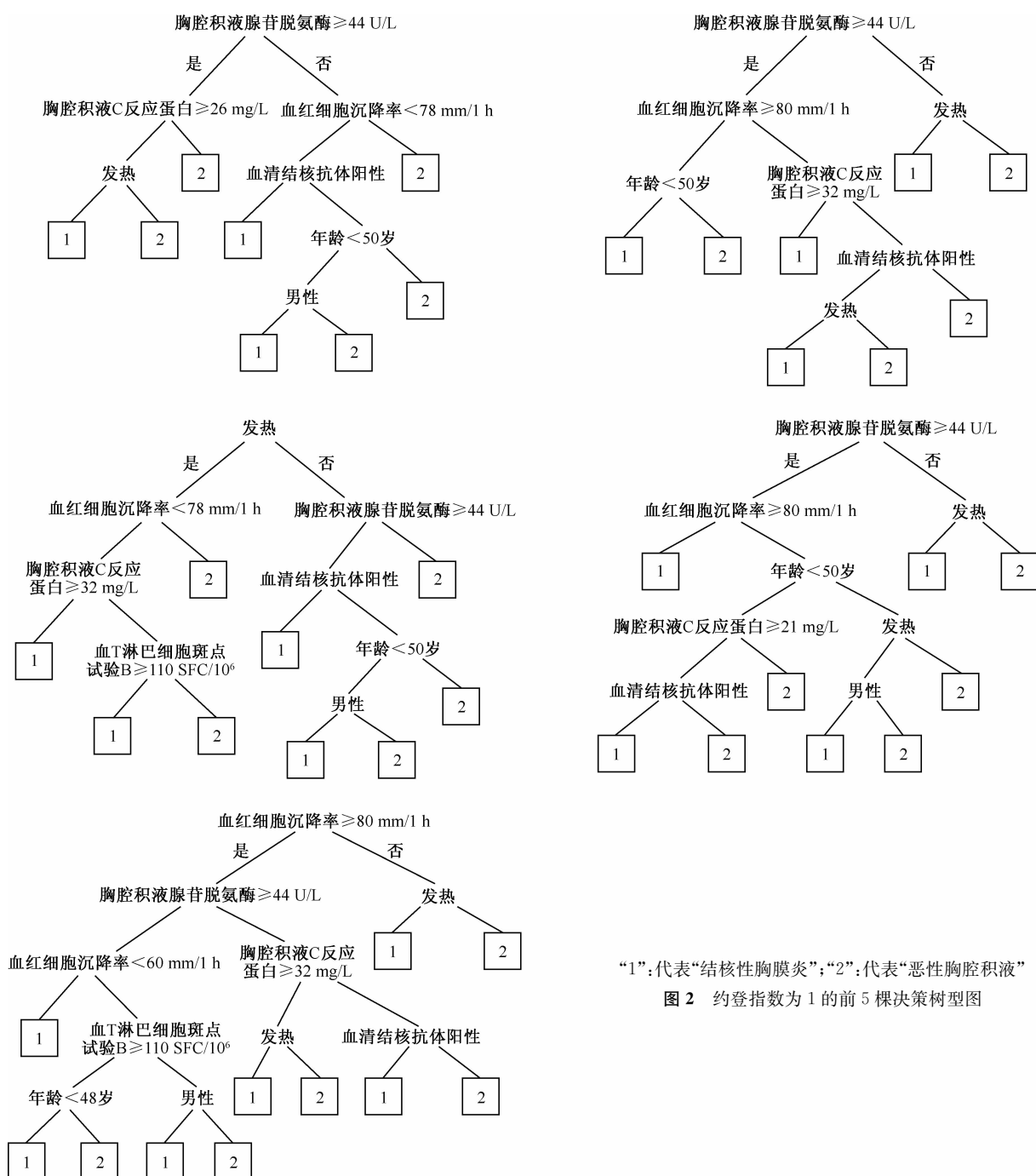
表 2 结核性胸膜炎决策树检测指标

检验指标	平均值(%)	标准差(%)
敏感度	98.14	3.13
假阴性率	1.86	3.13
特异度	93.64	7.44
假阳性率	6.36	7.44
约登指数	91.78	6.44
符合率	95.01	4.13
阳性预测值	94.17	4.38
阴性预测值	92.04	6.94
阳性似然比	11.80	6.94
阴性似然比	6.03	5.62

4. 决策树中各指标的贡献得分:决策树中各项指标的贡献得分是指各项指标在进行决策树诊断时所起的作用,即位于决策树顶端的指标,对诊断的作用较大。为探索对 TBP 诊断更为重要的因素,利用决策树分析将各指标的贡献得分进行加和与排序(图 3)。由此可以看出排序前 9 位的指标对诊断的贡献最大。

## 讨 论

TBP 是最常见的肺外结核,其诊断金标准是从病灶或胸腔积液中找到直接的细菌学证据。然而,目前临床能够拿到直接细菌学证据的诊断病例很少。目前,临床诊断主要依靠症状、体征及临床实验室检查,并凭借医生的经验综合判定,而关于综合性诊断研究尚少。张义宏等<sup>[7]</sup>利用年龄 $<37$ 岁、体温 $>37.8^\circ\text{C}$ 、胸腔积液腺苷脱氨酶 $>45$  U/L 或 $>30$  U/L、胸腔积液 C 反应蛋白 $>26$  mg/L 为截点,建立积分模型,利用积分模型作为判断结核性胸腔积液的依据。结果以积分模型分值 $\geq 7$ 分作为预测结核性胸腔积液的截点,其敏感度为 95.4%、特异度为 96.6%、准确性为 96.0%,取得了理想的结果。此外,采用构建临床诊断决策树的方法对难以获得直接证据的疾病进行诊断也是一种尝试。Porcel 等<sup>[8]</sup>采用决策树方法进行的 TBP 综合性诊断策略研究



“1”:代表“结核性胸膜炎”;“2”:代表“恶性胸腔积液”

图2 约登指数为1的前5棵决策树型图

中也证实,由他们制定的 TBP 诊断决策树对 TBP 与恶性胸腔积液鉴别诊断的敏感度为 92.2%、特异度为 98.3%、受试者工作特征曲线 (ROC 曲线) 下面积为 97.6%。临床诊断决策树与医生在进行临床诊断时的思维模式非常类似,因此可以作为一种诊断经验的实体形式,同时决策树又可以基于“大数据”而产生,从而提高了对疾病诊断的准确性。本研究以此为切入点,主要探讨这种研究方式的可能性及需要解决的问题。

本研究首先解决的问题是,利用临床的一些常

规数据来生成可以用于 TBP 诊断的临床诊断决策树。决策树的创建采用数据挖掘和机器学习的方法,因为机器学习的方法与临床思维有类似之处,容易为临床医生所接受。具体方法是以 R 语言中标准的 CART 决策树算法为核心,同时构建了包括数据的读取、数据清理、数据随机化分组、CART 决策树构建、检测指标的构建及相应结果输出的标准化流程,并在此基础上以 BioKepler 生物信息学分析平台为基础,构建 TBP 诊断决策树的工作流程。经对该工作流程进行千余次演算没有发现漏洞,且应

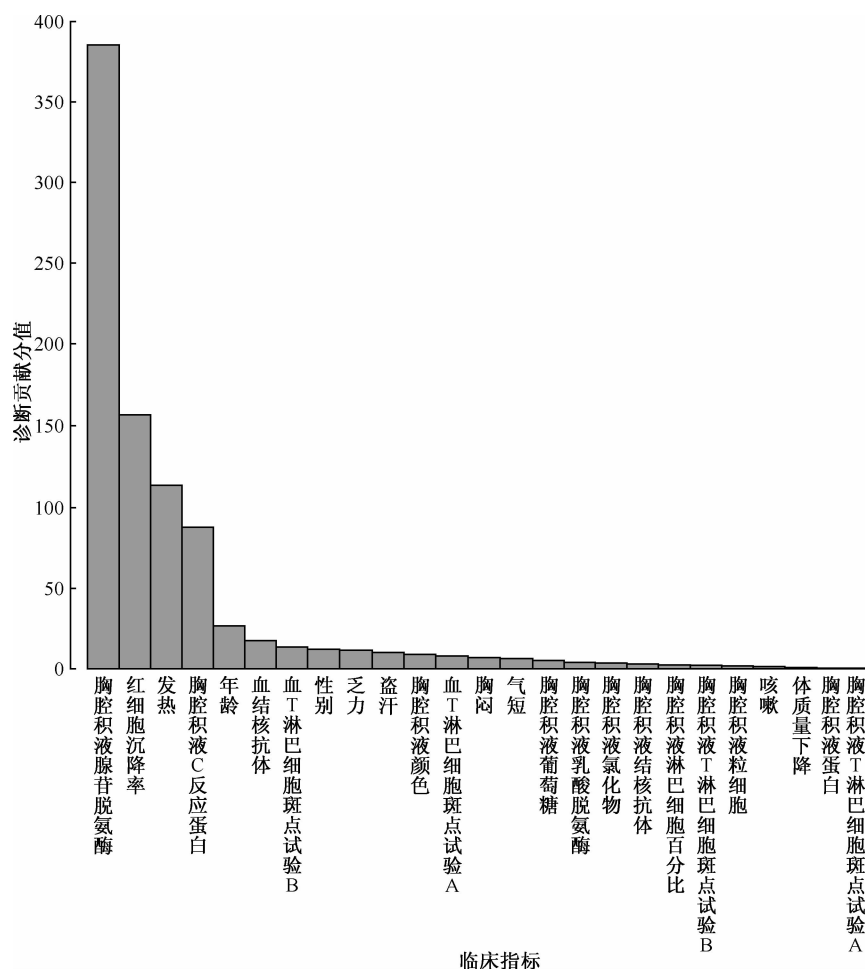


图 3 结核性胸膜炎各诊断因素贡献得分

用普通电脑可以在 30 s 左右计算出结果,因此初步解决了决策树在 TBP 诊断中的技术层面问题,为之后的研究奠定了基础。

第二个问题即所构建的临床诊断决策树能否达到一定程度的临床准确性。由于本次研究收集的患者例数较少,为提高所构建决策树诊断准确性,笔者应用已建立的工作流程,进行了 1000 次的模拟实验,并且对这 1000 次模拟实验的相应诊断实验检测指标进行了均值及标准差的计算。通过 1000 次循环模拟生成的决策树,对 TBP 与恶性胸腔积液鉴别诊断的敏感度均值为 98.14%,特异度均值为 93.64%。因此,初步显示应用临床诊断决策树的方法是可以达到一定的临床诊断准确性。

临床信息来源于多个渠道,如病史、查体和各种临床检测项目,本研究纳入了 25 项指标。哪些指标对诊断最有意义,直接关乎临床诊断结果。笔者首先对纳入的 25 项指标进行了组间比较,其中 16 项指标在两组之间差异具有统计学意义。而这 16 项指标在两种疾病鉴别时均有一定意义,是临床医师

对两种疾病进行鉴别诊断时需要考虑到的指标。而哪些指标对疾病鉴别诊断的意义更大,利用单因素分析不能明确,而利用决策树分析,通过指标的贡献得分则可以给出答案。笔者对模拟 1000 次生成的临床诊断决策树所包含的各项指标的贡献得分进行了排序。由于平均每个决策树需要 9 个指标作为决策因素,因此,笔者对贡献得分排名前 9 个指标进行了比较,发现其差异均存在统计学意义,因此,笔者从这两个方面验证了这 9 个指标对 TBP 的诊断价值更大。

通过决策树示意图可以看到,对 TBP 做出诊断的过程与临床医生的诊断过程非常相似,是综合了临床的各项指标给出的诊断,而不是某一项指标就能决定的。同时,决策树的生成可以并且应该基于大数据产生,这样生成的决策树会更加符合实际情况,对疾病诊断的准确性更高。陈蔼祥和陈智锋<sup>[9]</sup>利用机器学习方法构建的决策树用于结节病和肺结核的鉴别,准确性达到 96.15%。另外,利用决策

树方法进行疾病诊断时简便直观,只需要根据取得的临床数据来对照决策树,按照决策树分支一步步做出诊断。以图 1 为例简单说明诊断过程。首先看胸腔积液腺苷脱氨酶是否大于 44 U/L,根据是与否分别沿左右两个分支继续前进,直到符合某个诊断为止。同时,决策树所用到的指标为临床上容易获得的指标,即使是最基层医院也可以做到。因此,临床诊断决策树方法对于基层医院或经验不足的医生更有帮助。

由于本研究只是将决策树方法应用于 TBP 与恶性胸腔积液鉴别诊断的一次初步探索,目的是探讨利用决策树方法诊断 TBP 的可行性,还存在众多不足。(1)入选病历的局限性。本研究病历均为住院患者,这部分患者多年龄偏大,病情较重,以鉴别诊断病历较多。因此,患者的代表性较差。(2)本研究为回顾性研究,所收集的部分患者资料存在缺失问题,因此在统计学分析时会造成分析结果的不准确性。(3)样本量有限,因此不能形成较为稳定的临

床决策树, 尽管应用重复模拟的策略弥补了对决策树评价方面的缺陷, 但仍不能解决筛选稳定决策树进一步在临床进行前瞻性研究的问题。(4) 患者仅来源于同一家医疗机构, 致使所形成的决策树外推性不能得到有效的验证。(5) 对照组选择单一。仅选择了恶性胸腔积液作为对照组, 尽管该病为 TBP 最重要的鉴别诊断疾病之一, 但是只有对多种相关的疾病进行比较, 才能形成真正有效的临床诊断决策树。(6) 没有设立其他用于 TBP 诊断试验作为对照, 因此对于临床诊断决策树诊断 TBP 的实际效力还不能最终明确。尽管如此, 根据本研究得到的相应的研究结果, 我们认为已经达到了研究之初设定的研究目标, 并且为今后的临床研究提供了相关的研究线索及依据。

综上所述, 临床诊断决策树方法是诊断 TBP 的有效临床策略之一, 如进一步进行相关的循证医学研究, 有望解决目前 TBP 诊断手段缺乏和主观偏移性等问题。

## 参 考 文 献

- [1] Light RW. Update on tuberculous pleural effusion. *Respirology*, 2010, 15(3):451-458.
- [2] Thomas R, Lee YC. Causes and management of common benign pleural effusions. *Thorac Surg Clin*, 2013, 23(1):25-42.
- [3] 陈红兵, 周志红, 贺润年, 等. 荧光定量 PCR 技术在结核杆菌检测中的应用. *实用医学杂志*, 2008, 24(21):3765-3767.
- [4] 贝政平, 蔡映云. 内科疾病诊断标准. 2 版. 北京: 科学出版社, 2006.
- [5] Wang J, Crawl D, Altintas I. A framework for distributed data-parallel execution in the kepler scientific workflow system. *Procedia Computer Science*, 2012, 9(11):1620-1629.
- [6] 明鹤, 张应应. 基于 R 软件分析两组专家对五个葡萄酒样品的评分数据. *统计学与应用*, 2014, 3(4):133-140.
- [7] 张义宏, 张红梅, 沈景昊, 等. 用积分模型诊断结核性胸膜炎的价值研究. *中华医院感染学杂志*, 2014, 24(18):4512-4514.
- [8] Porcel JM, Alemán C, Bielsa S, et al. A decision tree for differentiating tuberculous from malignant pleural effusions. *Respir Med*, 2008, 102(8):1159-1164.
- [9] 陈蔼祥, 陈智锋. ADST: 用机器学习方法鉴别结节病和肺结核. *计算机科学*, 2014, 41(S1):103-109, 138.

(收稿日期: 2016-02-22)

(本文编辑: 李敬文)

## 第三届耐药结核病防控与诊治新进展研讨会征文通知

由中国防痨协会临床专业分会、《中国防痨杂志》期刊社、中国医疗保健国际交流促进会结核病防治分会、同济大学附属上海市肺科医院联合主办, 苏州市第五医院和江苏省疾病预防控制中心结核病防治研究所承办的“第三届耐药结核病防控与诊治新进展研讨会”拟于 2016 年 11 月 15 日—18 日(15 日为报到日, 18 日为撤离日)在苏州召开。本届论坛将邀请国内外著名结核病专家进行专题学术讲座, 并就耐药结核病的基础、诊断、治疗、预防等方面的最新研究成果及相关进展进行探讨。

1. 征文要求: (1) 稿件要求未在国内公开发行人物上发表(请在文题上方注明未公开发表, 未一稿多投); (2) 论著类稿件为全文+800 字左右的摘要, 摘要包括目的、方法、结果和结论, 也可仅提供符合上述要求的摘要; (3) 其他类型稿件为全文投稿; (4) 全文 4000 字以内, 编排顺序为: 题目、单位、邮编、姓名、中文摘要、正文、参考文献; (5) 本次会议征文不接收通过邮局邮寄的纸质版论文, 只接收 Word 版电子文件, 格式要求为题目 3 号黑体、正文 5 号宋体, 单倍行距;

(6) 请务必附第一作者与通信作者的通信地址、联系电话、手机、Email。

2. 征文发送注意事项: (1) 请通过 Email 发送至联系人邮箱, 邮件注明“耐药结核病会议征文”; (2) 截止日期: 2016 年 10 月 10 日。

3. 联系人: 李敬文, 手机: 13691010045; 电话(传真): 010-62257587; Email: lijwflzz@163.com。

入选论文将纳入会议《论文汇编》, 优秀论文将由大会学术委员会推荐刊登于《中国防痨杂志》或《结核病与肺部健康杂志》。参加会议者均可获得国家级继续医学教育学分证书。欢迎结核病及相关专业的临床医师及广大学者积极撰写会议征文并参加本次论坛。

中国防痨协会临床专业分会

《中国防痨杂志》期刊社

中国医疗保健国际交流促进会结核病防治分会

同济大学附属上海市肺科医院

2016 年 2 月