

DOI: 10.13288/j.11-2166/r.2017.17.011

## 临床研究

## 基于机器学习的卒中后抑郁影响因素分析

罗晓舟<sup>1</sup>, 温小鹏<sup>2</sup>, 何家扬<sup>3</sup>, 黄健婷<sup>1</sup>, 唐纯志<sup>1\*</sup>

(1. 广州中医药大学针灸康复临床医学院, 广东省广州市白云区机场路十二号, 510000; 2. 郑州市中心医院康复医学部; 3. 广州中医药大学热带医学研究所)

**摘要** 目的 通过机器学习判断脑卒中患者发生抑郁的影响因素。方法 从病历系统中提取符合纳入条件的 688 例脑卒中患者的病历资料, 包括年龄、性别、脉象、面色、舌质、舌苔、中医药干预方式、体重指数 (BMI)、血压、血糖、血甘油三酯、血总胆固醇、吸烟史、饮酒史、抑郁家族史、影像学卒中病灶部位及最终抑郁判断结果。采用单规则 (1R) 算法进行学习并判断提取信息中影响卒中后患者是否抑郁的危险因素, 继而将所搜集病例分为训练数据集 (500 例) 和测试数据集 (188 例), 并使用随机森林模型形成最优判别结果。结果 通过单规则算法得出脑卒中后是否抑郁最重要的影响因素为卒中病灶所在部位, 其中计算机推测卒中病灶位于额叶及颞叶者最易发生卒中后抑郁, 基底节、脑干、小脑、延髓、枕叶的病灶则不易引起抑郁, 其准确分类率达到 88.95% (612/688 例)。对前 500 例训练数据集进行随机森林模型判别, 其抑郁判断的正确率为 98.2%; 188 例测试集判断结果正确率达 99.47%; 将 688 例患者资料运用随机森林模型进行学习, 总的正确率为 98.84%。重要性测度结果显示, 病灶位置、中医药干预手段及抑郁家族史是脑卒中后是否发生抑郁最重要指标的前 3 位。结论 病灶位于额颞叶的脑卒中患者以及具有抑郁史的患者更容易发生卒中后抑郁。

**关键词** 脑卒中; 卒中后抑郁; 机器学习; 随机森林模型; 单规则算法; 集成学习

卒中后抑郁是指卒中后患者出现不同程度的情感低落、兴趣减退、悲观、厌世等症状, 且持续时间超过 2 周以上。医生对某种疾病的判断是基于医生多年的经验及多种体征、检查指标的综合判断, 而计算机能够在极短的时间里完成一个医生数年的学习积累。机器学习的核心思想是将已有信息转化为智能行为的算法<sup>[1]</sup>。随机森林模型是一种建立在决策树算法上的复合模型 (集成学习), 其主要特点在于随机、重抽样自举、竞争以及分组变量的随机性<sup>[2]</sup>。因为集成学习只需使用全体特征集中很小一个随机部分, 故其能处理相当庞大的数据, 尤其适用于医学数据这种庞大、多维、多特征的资料。本研究拟通过随机森林模型为主的机器学习方法, 对病历系统中 688 例卒中后抑郁患者资料进行分析, 探讨卒中后抑郁的危险因素及中医药干预手

段的可行性, 以总结出避免患者发生抑郁危险因素的方法或得到临床医生可以干预的措施指导。

## 1 资料与方法

### 1.1 病例信息来源及记录方法

病例资料来源于 2016 年 2 月至 2017 年 2 月广州中医药大学第一附属医院脑病科、针灸科及郑州市中心医院康复医学部神经康复科的住院部病历系统。按照时间顺序依次记录患者的基本资料、体征、四诊信息、治疗手段及抑郁判断结果, 包括: 年龄、性别、脉象、面色、舌质、舌苔、中医药干预方式、体重指数 (BMI)、血压、血糖、血甘油三酯、血总胆固醇、吸烟史、饮酒史、抑郁家族史、影像学卒中病灶部位及最终抑郁判断结果。抑郁结果在入院和出院时均进行判断, 除正常外皆记录为抑郁; 其他所有资料均以入院记录及首次病程记录为准。

### 1.2 诊断标准

参照《各类脑血管疾病诊断要点》脑卒中诊

基金项目: 国家重点基础研究发展计划 (“973” 计划) (2012CB518504)

\* 通讯作者: jordan664@163.com

断标准<sup>[3]</sup>。

### 1.3 各指标分级界定标准

1) 年龄: 本研究设 <45 岁为青年, 45 ~ 59 岁为中年, >60 岁为老年。2) BMI: 按照中国肥胖工作组 2003 年制订的标准<sup>[4]</sup>, BMI <18.5 为偏轻, 18.5 ≤ BMI <24 为正常, 24 ≤ BMI <28 为超重, BMI ≥ 28 为肥胖。3) 血压: 收缩压 90 ~ 140 mmHg、舒张压 60 ~ 90 mmHg 为正常; 低于正常标准为低血压; 收缩压 140 ~ 159 mmHg、舒张压 90 ~ 99 mmHg 为 1 级高血压; 收缩压 160 ~ 179 mmHg、舒张压 100 ~ 109 mmHg 为 2 级高血压; 收缩压 ≥ 180 mmHg、舒张压 ≥ 110 mmHg 为 3 级高血压。4) 血糖: 空腹血糖 3.9 ~ 6.1 mmol/L 为正常, 高于此标准判定为偏高。5) 血甘油三酯: 空腹全血甘油三酯值 0.22 ~ 1.65 mmol/L 范围内为正常, 高于此标准为偏高。6) 血总胆固醇: 空腹全血总胆固醇低于 5.17 mmol/L 为正常, 高于此标准判定为偏高。7) 抑郁家族史: 采用医患谈话的方式对患者自述或家属代述情况进行记录, 结果分“有”或者“无”。8) 卒中病灶定位: 选用患者入院后或入院前 1 个月内的头颅 CT 或 MRI 结果报告为准。9) 抑郁判定标准: 采用汉密顿抑郁评分量表 (HAMD) 评定, 总分 <8 分为正常, ≥8 分者判断为抑郁<sup>[5]</sup>。10) 住院总时间设定: 考虑到医保政策、既往患者平均住院时间及各种因素的影响, 课题组成员商议后将住院总时间定为两周。即若患者入院后 HAMD 评分已达抑郁者直接判断为“抑郁”, 未达抑郁者, 在住院 2 周后再行评定作为结果。

### 1.4 纳入标准

符合上述诊断标准; 符合各指标分级界定标准; 年龄 35 ~ 75 岁; 配合完成治疗者。

### 1.5 排除标准

合并其他系统严重疾病; 意识障碍无法完成量表评估; 资料不齐全; 疗效判定标准不同; 曾因本病行手术治疗; 住院时间 <2 周。

### 1.6 统计学方法

采用 Rx64 2.15.3 程序包中的 RWeka 程序包及 random Forest 程序包对纳入记录的患者信息进行机器学习, 并采用 STATA 14.0 软件进行计算。首先采用单规则 (1R) 算法进行学习并判断提取信息中影响卒中后患者是否抑郁的危险因素, 继而将所搜集病例按照一定比例分为训练数据集和测试数据集并使用随机森林模型形成最优判别结果。

## 2 结果

### 2.1 患者一般情况

最终纳入患者 688 例。年龄分布: 青年 93 例, 中年 391 例, 老年 204 例; 性别分布: 男性 301 例, 女性 387 例; 脉象共录得 8 种: 沉脉 8 例, 滑脉 235 例, 平脉 87 例, 濡脉 284 例, 涩脉 5 例, 数脉 3 例, 细脉 22 例, 弦脉 44 例; 面色: 色白 57 例, 色黑 39 例, 色红 408 例, 色黄 100 例, 晦暗 37 例, 正常 (红黄隐隐, 明润含蓄) 47 例; 舌质: 淡 178 例, 淡红 406 例, 舌质老 12 例, 舌质嫩 41 例, 舌胖大 49 例, 瘦舌 2 例; 舌苔: 白苔 381 例, 薄白苔 2 例, 剥落苔 3 例, 黄腻苔 227 例, 少苔 75 例; 中医药干预: 针药结合干预 288 例, 单纯针灸干预 82 例, 单纯中药干预 318 例; BMI: 偏轻 1 例, 正常 265 例, 超重 252 例, 肥胖 170 例; 血压: 低血压 16 例, 正常 43 例, 1 级高血压 315 例, 2 级高血压 207 例, 3 级高血压 107 例; 血糖: 正常 326 例, 偏高 362 例; 血甘油三酯: 正常 144 例, 偏高 544 例; 血总胆固醇: 正常 294 例, 偏高 394 例; 吸烟: 不吸烟 495 例, 吸烟 193 例; 饮酒: 不饮酒 194 例, 饮酒 494 例; 病灶定位: 额叶 19 例, 基底节 88 例, 脑干 202 例, 颞叶 72 例, 小脑 50 例, 延髓 69 例, 枕叶 188 例; 抑郁: 不抑郁 521 例, 抑郁 167 例。

### 2.2 单规则 (1R) 算法学习结果

通过单规则算法得出卒中后是否抑郁最重要的影响因素为卒中病灶所在部位, 其中计算机推测卒中病灶位于额叶及颞叶者最易发生卒中后抑郁, 基底节、脑干、小脑、延髓、枕叶的病灶则不易引起抑郁, 其准确分类率达到 688 例中的 612 例 (正确分类率达 88.95%)。结果中没有 1 例是本来“不抑郁”的被错分到“抑郁”类; 而有 76 例原本是“抑郁”的被错分到了“不抑郁”类。

综上, 单规则算法认为卒中后抑郁的发生最重要的因素是卒中病灶的位置。考虑到卒中后抑郁可能导致患者轻生等严重后果, 不得不重视其发生原因及危险因素, 且该算法本身具有局限性 (仅考虑一项最重要的规则), 因此进一步采用复合模型的随机森林算法来继续探讨。

### 2.3 训练及测试集学习效果随机森林模型评价结果

首先通过计算机生成随机数字, 对 688 例患者进行随机排序, 将 1 ~ 500 例作为训练数据集, 501 ~ 688 例作为测试数据集进行模型评估。

#### 2.3.1 训练集最终结果 将 500 例训练集数据进

行随机森林模型判别之后,得到的结果在“不抑郁”的 375 例中被错分 4 例,错误分类率为 1.07%;在“抑郁”的 125 例中被错分 5 例,错误分类率为 4.00%。总的正确率为 98.2%,相较于单规则模型得到了极大地提升。出包错误率(out-of-bag error rate)为 1.8%,这是一个对于该模型未来性能的合理估计。总共建立了 500 棵决策树。

2.3.2 测试集最终结果 将 188 例作为测试集测试前文训练所建立模型的效果判断。结果显示,该 188 例患者中仅有 1 例本应是“抑郁”患者被错误地划分到了“不抑郁”组,正确率高达 99.47%。

#### 2.4 全部患者随机森林模型预测学习结果

为探讨患者体征、四诊资料等对卒中后患者是否发生抑郁的影响,进一步将所有 688 例患者资料运用随机森林模型进行学习。将 688 例病例数据进行随机森林模型预测之后,得到的结果在“不抑郁”的 521 例中被错分 5 例,错误分类率为 0.096%;在“抑郁”的 167 例中被错分 3 例,错误分类率为 1.80%。总的正确率为 98.84%,相较于仅用 500 例训练数据集的结果稍有提升,提示随着病例数据量的增加,本模型的准确度还将继续增加。出包错误率为 1.16%,较前减少。总共建立了 500 棵决策树。随机森林模型的结果中可将每例病例 500 棵决策树的判断比例进行调用,将前 6 例病例资料展示如表 1。

表 1 前 6 例脑卒中病例随机森林模型决策树判断比例

| 病历资料序号 | 不抑郁         | 抑郁        |
|--------|-------------|-----------|
| 1      | 0           | 1         |
| 2      | 0.176829268 | 0.8231707 |
| 3      | 0           | 1         |
| 4      | 0.145077720 | 0.8549223 |
| 5      | 0.005405405 | 0.9945946 |
| 6      | 0.020100503 | 0.9798995 |

由表 1 可知,对于第 1 例患者的资料模型学习后,500 棵决策树全部将其判断为了“抑郁”;而对第 2 例患者决策树之间出现了不同意见,500 棵决策树中 17.7% 将其判断为“不抑郁”,82.3% 将其判断为“抑郁”,最终计算机以少数服从多数的方法将其判断为了“抑郁”。其他病例均同此例判断。计算结果中表明随机森林的错判率,“不抑郁”类的预测效果好于整体及“抑郁”类的判断,而当决策树达到 320 棵左右以后,各类错判率基本达到稳定。最后将上述模型运用于全部资料的预测分类,得到的结果显示全部 688 例资料中仅有 1 例本为“不抑郁”的患者被错分为“抑郁”,正确率

高达 99.85%,相较于仅用 500 例作为训练集的正确率再次有所提高。

#### 2.5 卒中后抑郁影响因素分析

各因素对于结果预测精度影响变化方面,通过计算输入变量的重要性测度得出:病灶位置、中医药干预手段及抑郁家族史是患者是否可能并发卒中后抑郁最重要指标的前 3 位,而最不重要者为患者的血甘油三酯是否正常。病灶位置已在单规则算法学习结果得出,中医药干预手段及抑郁家族史患病例数通过卡方检验,结果见表 2。由表 2 可知,不同干预手段和有无家族抑郁史患者发生卒中后抑郁的分布不同( $P < 0.05$ );其中单纯中药干预及具有抑郁家族史的患者发生卒中后抑郁的机会较高。采用卡方检验对单纯中药和单纯针灸干预的两组进行检验,两组患者患有抑郁的比例差异具有统计学意义( $\chi^2 = 23.3373, P < 0.05$ ),提示卒中后患者采用单纯针灸干预发生抑郁的情况少于单纯中药干预。

表 2 脑卒中患者是否抑郁在中医药干预手段及抑郁家族史方面分布情况(例)

| 患者抑郁情况     | 中医药干预手段  |    |     | 抑郁家族史   |     |
|------------|----------|----|-----|---------|-----|
|            | 针药结合     | 针灸 | 中药  | 有       | 无   |
| 不抑郁        | 288      | 67 | 166 | 57      | 464 |
| 抑郁         | 0        | 15 | 152 | 141     | 25  |
| $\chi^2$ 值 | 189.6578 |    |     | 336.034 |     |
| $P$ 值      | <0.005   |    |     | <0.005  |     |

#### 3 讨论

单规则(1R)算法学习所得到的结果显示,卒中患者发生卒中的脑区是其卒中后是否发生抑郁最重要的影响因素,额叶及颞叶的病灶更有机会引起患者卒中后抑郁;而颞叶受损引起抑郁的报道目前还不多见,颞叶主管人体听觉语言中枢、记忆、嗅觉及情感等,但其具体影响人体精神状态的机制尚不明确。本研究的主要目的虽然并不是探讨额叶、颞叶对人体精神活动的影响,但其结果已从侧面证实了其与精神活动的关联性,而本部分研究最重要的意义是将两者的联系具化地提出来,提示临床医生遇到额颞叶受损的患者时应及时采取相应措施,例如嘱咐其家属的陪同与开导,进而防范该类患者发生卒中后抑郁导致更加严重的后果如患者轻生等。本研究机器学习结果正确率相较于临床误判率(15%~40%<sup>[8]</sup>),并不能说其结果太差。

随机森林模型在本研究中首先将所有患者资料分为了训练集(训练模型)与测试集(测试模型),再采用训练集所训练处的随机森林模型对测

试集进行判别, 其对测试数据集的判断准确率高达 99.47%。这虽然能够证明该模型判断的准确率, 但无法说明影响患者卒中后是否发生抑郁的关键因素。因此接下来将全部资料交予计算机进行预测, 其判断准确率达到 98.84%。相较于仅用 500 例资料训练出的模型再次有所提高, 这也提示我们随着病例资料的丰富, 计算机的识别准确率将继续提高。最后, 我们调用模型所给出的参数明确了卒中后患者抑郁发生率的影响因素, 其前 3 位分别是卒中病灶位置、中医药干预手段以及卒中患者是否具有抑郁家族史。虽然本模型不能直接提示我们各因素不同水平对最终结果的影响大小, 但通过卡方检验可得出以下结论: 1) 病灶位于额颞叶的卒中后抑郁发生率明显高于病灶位于其他位置的患者; 2) 单纯采用中药干预以及家族中具有抑郁史的患者发生卒中后抑郁的概率也明显较高。

综上所述, 随着数据量呈指数级的增长, 人工智能以及其包含的机器学习、深度学习等方法不断完善, 我们虽不能在此断言机器终将取代人类, 但该方法的普及应用必将极大地减少临床医生在诊断和决策时所耗费的时间。本研究的不足之处在于

囿于各种因素, 且患者资料来自不止一家医院, 在记录中医药干预手段时分类相对较泛, 并未对每一例患者汤剂处方、针灸穴位及操作手法等记录。但通过本研究结论已经可以明确地得出, 中医药干预特别是针药结合的干预是一个有效减少卒中后患者发生抑郁的手段。

## 参考文献

- [1] BRETT LANTZ. 机器学习与 R 语言 [M]. 北京: 机械工业出版社, 2015: 82-112.
- [2] 薛薇. R 语言与数据挖掘 [M]. 北京: 中国人民大学出版社, 2016: 194-201.
- [3] 中华神经科学会. 各类脑血管疾病诊断要点 [J]. 中华神经科杂志, 1996, 29(6): 379.
- [4] 中国肥胖问题工作组. 中国成人超重和肥胖症预防与控制指南(节录) [J]. 营养学报, 2004(1): 1-4.
- [5] BECH P, KASTRUP M, RAFAELSEN OJ. Mini-compendium of rating scales for states of anxiety depression mania schizophrenia with corresponding DSM-III syndromes [J]. Acta Psychiatr Scand Suppl, 1986, 326: 1-37.
- [6] 刘锋, 董丽华, 龚守良, 等. 从临床误诊率看诊断学教学改革必要性 [J]. 中国高等医学教育, 2002, 38(6): 38-39.

## Analysis of the Influencing Factors of Post-stroke Depression: Based on Machine Learning

LUO Xiaozhou<sup>1</sup>, WEN Xiaopeng<sup>2</sup>, HE Jiayang<sup>3</sup>, HUANG Jianting<sup>1</sup>, TANG Chunzhi<sup>1</sup>

(1. Acupuncture & Rehabilitation Clinical Medical School, Guangzhou University of Chinese Medicine, Guangzhou, 510000; 2. Rehabilitation Medicine Department, Zhengzhou Central Hospital; 3. Institute of Tropical Medicine, Guangzhou University of Chinese Medicine)

**ABSTRACT Objective** To determine the influencing factors of post-stroke depression by machine learning.

**Methods** Stroke patients' medical records (688 cases eligible) were extracted from record system, including age, gender, pulse manifestation, complexion, tongue quality, fur, Chinese medicine intervention, body mass index (BMI), blood pressure, blood glucose, blood triglyceride, blood total cholesterol, smoking history, drinking history, depression family history, stroke lesion site in imaging, as well as the final depression judgment. Single rule algorithm (1R) was adopted to learn. The risk factors influencing post-stroke patients' depression in extracted information were determined. Then the cases collected were divided into the training dataset (500 cases) and the test dataset (188 cases). Optimal discriminant results were obtained by random forest model. **Results** Single rule algorithm showed that the most important influencing factor of post-stroke depression was stroke lesion site. By computer speculation, stroke lesions in the frontal and temporal lobes were most prone to post-stroke depression. Basal ganglia, brain stem, cerebellum, medulla oblongata and occipital lobe lesions were less likely to cause depression. The accurate classification rate could amount to 88.95% (612/688 cases). Random forest model determination was made in the former 500 cases in the training dataset. The total correct rate of determining depression was 98.2%. The total correct rate of determination in 188 cases of the test dataset was 99.47%. Six hundred and eighty-eight patients' data were learnt by random forest model. The total correct rate was 98.84%. The importance measure results showed that top 3 important indexes of post-stroke depression were lesion site, Chinese medicine intervention and depression family history. **Conclusion** Patients with lesions in the frontal & temporal lobes and depression family history were most prone to post-stroke depression.

**Keywords** stroke; post-stroke depression; machine learning; random forest; single rule algorithm; ensemble learning

(收稿日期: 2017-04-05; 修回日期: 2017-05-29)

编辑: 崔京艳