

DOI:10.13288/j.11-2166/r.2022.07.006

顽固性高血压发生主要不良心血管事件患者的 中医四诊信息可解释性研究 ——基于随机森林规则提取方法

向兴华^{1,2}, 彭叶辉^{1✉}, 杨伟^{2✉}, 刘剑³, 刘大胜², 纪鑫毓², 雷舒杨²,
谢飞彪^{1,2}, 杨纪雯², 王丽颖², 韩学杰², 商洪才⁴

1. 湖南科技大学数学与计算科学学院, 湖南省湘潭市雨湖区桃园路, 411201; 2. 中国中医科学院中医临床基础医学研究所;
3. 中国中医科学院西苑医院; 4. 北京中医药大学东直门医院

[摘要] **目的** 基于随机森林规则提取方法探讨顽固性高血压(RH)发生主要不良心血管事件(MACE)患者的临床信息分类规则及特征。**方法** 采集住院治疗的2093例RH患者的电子病历信息,其中RH患者807例,RH+MACE患者1286例。收集患者的个人基本信息(包括性别、年龄、既往史、家族史、个人史、传染病史)、中医四诊信息、西医诊断信息数据。统计前期研究发现的与RH患者发生MACE密切相关的140个指标的分布情况,并基于密切相关指标运用随机森林规则提取方法分析RH患者发生MACE的影响特征;对数据进行结构化处理并基于随机森林训练构造多个去相关决策树,生成丰富的决策规则,再采用规则提取方法从规则中提取、度量、修剪、选择规则,利用序贯覆盖法筛选简化的分类规则集,并定义规则的可解释性度量。**结果** 2093例患者中医四诊信息频次前5位的密切相关指标分别为弦脉1101次(52.6%)、苔腻947次(45.25%)、睡眠欠佳870次(41.57%)、面色少华784次(37.46%);患者基本信息及西医诊断信息频次前5位分别为心率正常1765次(84.33%)、既往高血压病1601次(76.49%)、年龄大于60岁1547次(73.91%)、舒张压正常1428次(68.23%)、高血压3级1151次(54.99%)。RH患者发生MACE影响特征的可解释性分析结果显示,RH+MACE且误差 <0.05 的规则有3条,规则1:咳嗽c('1') & 老舌c('1') & 头晕c('0') & 胸闷c('0');规则2:恶心c('0') & 颈动脉硬化c('1') & 弦脉c('1');规则3:憋喘c('1')。3条规则分类准确率均大于95%(误差小于0.05),并具有一定的临床可解释性。分类规则集在准确性与可解释性方面的统计性能分别为灵敏度0.85、特异度0.73、精度0.83、受试者工作特征曲线下面积(AUC值)0.80以及可解释性值0.96。**结论** 咳嗽、老舌、颈动脉硬化、脉弦、憋喘且无头晕、胸闷、恶心症状为RH患者发生MACE的核心临床信息分类要素。

[关键词] 顽固性高血压; 主要不良心血管事件; 随机森林规则; 可解释性研究

顽固性高血压(resistant hypertension, RH)指在改善生活方式的基础上,使用可耐受、足够剂量且合理的3种降压药(包括利尿剂)治疗至少4周后,血压值仍在目标水平之上,或至少需要4种药物才能使血压达到标准值^[1]。RH具有病情复杂、病程漫长,且多涉及心、脑、肾等多脏器并发症的

特点,极易引发主要不良心血管事件(major adverse cardiovascular events, MACE)的发生,包括心血管死亡、心肌梗死与脑卒中,严重影响患者的预后及生存质量^[2]。

中医学认为,RH主要因疾病日久,脏腑气血阴阳失调,同时风、火、痰、瘀等病理因素相互兼夹,导致虚实夹杂,故易生变证,预后较差^[3]。但因RH病机变化多端,临床辨证施治的精准性仍存在一定困难。近年来,研究者利用真实世界中医电子病历(electronic medical record, EMR)观察性

基金项目:国家重点研发计划(2017YFC1700400,2017YFC1700406-2);国家自然科学基金(81774454,61771491)

✉ 通讯作者:彭叶辉, pengyehui@hnust.edu.cn; 杨伟, yangyxq@ruc.edu.cn

数据,通过统计机器学习发现潜在有用的规律或知识,形成对疾病的预测或分类,可提高诊断准确率^[4-6]。前期已有研究采用决策树、随机森林、GBM、XGBOOST等模型对RH患者是否合并MACE的影响特征及关键“症-药”规律进行分析,并得出了模型的分类性能及各个影响特征的重要性程度^[7],但尚无法确定这些特征取值组合是如何评估RH患者是否合并MACE的。本研究根据随机森林规则提取方法的可解释性框架技术^[8-11],优化随机森林生成的规则集,显示分类过程中可解释的规则和模式,进一步探究RH患者发生MACE的影响因素及其取值组合对发生MACE的影响,为临床医生采用合理且有效的方案预防RH患者发生MACE提供依据。

1 临床资料

1.1 研究对象

从中国中医科学院西苑医院电子病历管理系统(EMR系统)选取2015年1月1日至2019年11月30日住院治疗的RH患者与RH+MACE患者共2093例,其中男性977例(46.7%),女性1116例(53.3%);平均年龄(68.57±14.06)岁;RH患者807例,RH+MACE患者1286例。本信息采集严格按照中国中医科学院西苑医院保密信息采集,对患者的敏感信息进行脱敏,严格遵循伦理原则要求。

1.2 诊断标准

RH与MACE诊断标准均参照《中国高血压防治指南2010》^[12]和《中国高血压防治指南(2018年修订版)》^[13]制定。

1.3 纳入标准

1) RH人群是指符合RH诊断,且中医、西医诊断名与既往史中不含有MACE相关疾病(心、脑、肾等)诊断的患者;RH+MACE人群是指符合RH诊断,且中医、西医诊断名与既往史中含有MACE相关疾病诊断的患者。2) 患者一次就诊记录的西药药品包含至少3种降压药物(至少包含1种利尿剂),开药时间间隔大于4周且就诊次数在2次以上,或西药记录中至少有一次包含4种及以上降压药物。3) 患者信息记录中包含完整的住院病历信息(性别、年龄、主诉、刻下症、既往史、家族史、个人史、传染病史、望诊、闻诊、问诊、舌象、脉象信息)。

2 方法

2.1 数据处理及特征说明

基于EMR系统采集的主诉、刻下症等相关中

医四诊信息多以文本非结构化的形式为主,需要进行数据预处理,具体包括术语分词、术语规范、术语命名实体识别及术语数据清洗等自然语言处理过程。主要采用以下方式:1) 对分词后的症状数据采用频数特征规范化以及自然语言理解^[14]两种方法相结合的处理方式;2) 对数据中缺失数据、冗余数据、噪声数据进行规范清洗。另外,定义结局特征为RH患者人群是否合并MACE,记为“RH+MACE”,若发生取值为1,不发生取值为0。

2.2 数据分析

前期研究^[7]筛选出140个与RH患者发生MACE密切相关的指标。本研究对2093例RH患者中医四诊信息、基本信息及西医诊断信息进行频数统计,观察140个密切相关指标分布情况,采用频数(%)表示。对140个密切相关指标进行随机森林规则提取方法分析,在建立随机森林模型的过程中对决策树进行多次迭代,且发现决策树数目大于200后,随机森林模型误差趋于稳定。按照可解释性框架建立决策树模型、随机森林模型,并对随机森林模型进行规则提取,对提取的规则编程优化,构建简化规则集并计算准确性与可解释性的度量。数据分析采用R 3.6.2统计软件,并使用R软件包rpart、randomForest、inTrees及caret处理程序。

随机森林规则提取方法的可解释性框架分析是指从具有众多无关或冗余规则的规则集中获取精炼有效的规则。具体步骤:1) 生成决策规则:构建多棵决策树组成的随机森林模型,每棵决策树从根节点到叶节点产生的IF-THEN规则^[15]可以表示为:IF<条件C>THEN<结局T>,记为{C⇒T},C是由特征值对组成的规则条件,T称为规则条件对应的结局或目标;2) 提取规则:为了保证数据的完整性,提取随机森林模型产生的规则条件,并以原始训练集为规则条件重新分配结局;3) 度量规则:引用三个度量指标来度量规则的统计特性^[9],即频率、误差和长度;4) 修剪规则:采用Leave-One-Out修剪法^[9]对规则中的条件进行修剪;5) 选择规则:将修剪后的规则集通过特征选择方法(即将规则条件转化成二分类特征后),从中选择一组相关但不冗余的特征子集^[16],具体的特征选择算法采用引导式正则化随机森林^[17];6) 简化规则集:采用序贯覆盖法^[18-19]总结一个简化的最终规则集,该算法在训练数据中进行多次迭代,目的是建立一个由最佳规则(最佳规则定义为训练数据误差最小的规则。如果两条规则误差一样,则选择频率较高、长度较小的规则)构成的有序规则列

表; 7) 定义可解释性指标:

$$Interpretability(N_{INT}) = 1 - \frac{\alpha N_{FEA} + \beta N_{COV} + \gamma N_{CNT}}{\alpha + \beta + \gamma}$$

N_{FEA} 、 N_{COV} 和 N_{CNT} 分别定义为:

$$N_{FEA} = \frac{1}{rule_{selected}} \sum_{i=1}^Z \frac{active_{fea}^i - 1}{input_{fea}^i - 1},$$

$$N_{COV} = \frac{1}{rule_{selected}} \left(1 - \sum_{i=1}^{ZZ} \frac{rule_{match}^i - 1}{count_{data} - 1} \right),$$

$$N_{CNT} = \frac{rule_{selected} - 1}{rule_{count} - 1}$$

其中 N_{FEA} 、 N_{COV} 和 N_{CNT} 分别表示规则的平均复杂度、每条规则的覆盖范围以及规则选择率; α 、 β 和 γ 分别表示各指标的权重, 其可根据研究者的背景确定, 此处设定为 1 即权重相等^[20], 这 3 项指标分别反映了规则集的简洁性、稳定性以及预测性。

3 结果

3.1 2093 例 RH 患者 140 个密切相关指标分布情况

2093 例患者中医四诊信息频次排名前 10 的指标分别为: 弦脉 1101 次 (52.60%)、苔膩 947 次 (45.25%)、睡眠欠佳 870 次 (41.57%)、面色少华 784 次 (37.46%)、胸闷 775 次 (37.03%)、两目乏神 773 次 (36.93%)、老舌 763 次 (36.45%)、苔薄 735 次 (35.12%)、乏力 733 次 (35.02%)、头晕 713 次 (34.07%), 频率大于 10% 的指标分布情况具体见表 1。

表 1 2093 例顽固性高血压患者频率 > 10% 的
中医四诊信息分布情况

指标	频次	频率/%	指标	频次	频率/%
弦脉	1101	52.60	精神欠佳	620	29.62
苔膩	947	45.25	沉脉	440	21.02
睡眠欠佳	870	41.57	心悸	440	21.02
面色少华	784	37.46	动作迟缓	391	18.68
胸闷	775	37.03	大便秘结	389	18.59
两目乏神	773	36.93	咯痰	384	18.35
老舌	763	36.45	憋喘	364	17.39
苔薄	735	35.12	喘促	360	17.20
乏力	733	35.02	气短	330	15.77
头晕	713	34.07	憋气	271	12.95
食纳欠佳	690	32.97	胸痛	260	12.42
下肢水肿	677	32.35	痰色白	246	11.75
苔黄	654	31.25	痰质黏	236	11.28
咳嗽	627	29.96	恶心	233	11.13

患者基本信息 (包括性别、年龄、既往史、家族史、个人史、传染病史) 及西医诊断信息频次排名前 10 的指标分别为: 心率正常 1765 次

(84.33%)、既往高血压病 1601 次 (76.49%)、年龄大于 60 岁 1547 次 (73.91%)、舒张压正常 1428 次 (68.23%)、高血压 3 级 1151 次 (54.99%)、男性 977 次 (46.68%)、既往糖尿病 687 次 (32.82%)、收缩压正常 587 次 (28.05%)、糖尿病血糖稳定 548 次 (26.18%)、既往高脂血症 489 次 (23.36%), 频率大于 10% 的指标分布情况具体见表 2。

表 2 2093 例顽固性高血压患者频率 > 10% 的基本信息
及西医诊断信息分布情况

指标	频次	频率/%	指标	频次	频率/%
心率正常	1765	84.33	低钾血症	398	19.02
既往高血压病	1601	76.49	高血压病家族史	344	16.44
年龄大于 60 岁	1547	73.91	既往脑梗死	327	15.62
舒张压正常	1428	68.23	细菌性肺炎	305	14.57
高血压 3 级	1151	54.99	颈动脉硬化	284	13.57
男性	977	46.68	肢体动脉硬化	275	13.14
既往糖尿病	687	32.82	低蛋白血症	259	12.37
收缩压正常	587	28.05	慢性肾功能不全	243	11.61
糖尿病血糖稳定	548	26.18	糖尿病性周围神	242	11.56
既往高脂血症	489	23.36	经病		
脂肪肝	463	22.12	既往心律失常	230	10.99
反流性食管炎	426	20.35	糖尿病伴血糖控	230	10.99
高尿酸血症	401	19.16	制不佳		

3.2 RH 患者发生 MACE 影响特征的可解释性分析

采用随机森林规则提取方法的可解释性框架对 140 个密切相关指标即影响特征进行可解释性分析, 重点考虑 RH 患者中医四诊信息及其有效规则对 RH+MACE 分类的影响。将 2093 例 RH 患者按 7:3 划分训练集和测试集, 训练集 1482 例, 其中 RH+MACE 患者 919 例; 测试集 611 例, 其中 RH+MACE 患者 367 例。

3.2.1 RH 患者发生 MACE 的规则度量及修剪 对训练集 1482 例 RH 患者以是否合并 MACE 建立随机森林模型, 并结合规则提取方法提取出 5336 条分类 RH+MACE 的规则 (影响特征的组合), 其中存在冗余和相关的规则, 为了便于临床解释, 我们对提取的规则进行去重, 并采用频率、误差和长度三个指标进行度量, 然后对每条规则中的影响特征进行修剪。如提取的规则集有规则 1: {年龄大于 60 岁('1') & 既往高血压('1') & 憋喘('1') $\Rightarrow$ RH + MACE}, 规则 2: {年龄大于 60 岁('1') & 既往高血压('1') & 憋喘('1') & 头痛('0') $\Rightarrow$ RH + MACE}, 规则 3: {年龄大于 60 岁('1') & 既往高血压('1') & 喘憋('1') $\Rightarrow$ RH + MACE}。去除重复规则 3 并对规则 1、

规则 2 进行度量, 修剪删除规则 2 中的头痛症状, 剩余影响特征值对误差与原始规则 2 误差相比较, 若误差值在预先设置的范围内, 则删除头痛, 反之则保留。通过对提取的 5336 条规则去重、度量以及修剪, 保留了 2817 条分类规则, 规则集修剪前后的规则长度、规则频率以及规则误差的变化如图 1 所示。修剪前后可明显看出规则长度减小, 规则频率增大, 规则集误差减小。

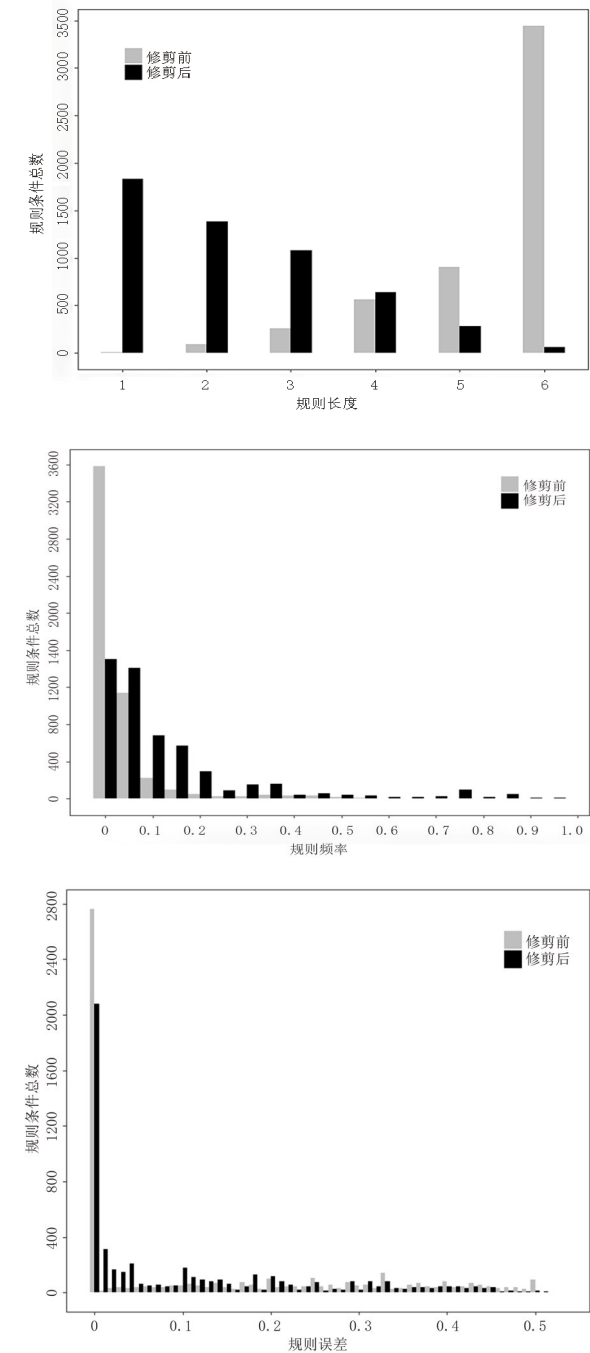


图 1 顽固性高血压患者发生主要不良心血管事件随机森林规则提取方法规则修剪前后情况

3.2.2 RH 患者发生 MACE 的规则选择及简化 基于 Gini 系数的平均下降统计量对随机森林模型所有影响特征进行重要性排序, 排名前 30 位的影响特征包括中医临床症状 28 个、舌象 1 个、脉象 1 个, 但无法解释这些特征取值组合如何影响结局 RH+MACE 的分类, 故考虑可解释性框架技术对以上影响特征构建可解释性的简化规则集。

根据以上分析, 发现修剪规则后仍存在大量的规则, 且根据频率、误差对规则进行排序, 排在前面的规则可能彼此相似即存在冗余, 故采用引导式正则化随机森林特征选择算法, 把每条规则条件转化成二分类特征后, 从中选择一组相关但不冗余的特征子集。为了最终形成可解释性分类工具, 采用序贯覆盖法^[18-19]总结一个简化的最终规则集, 如表 3 所示。结果显示, 应用于 RH+MACE 患者分类的规则有 5 条, 其中规则 1、3 和 6 的分类准确率都达到了 95% 以上 (误差小于 0.05), 并具有一定的临床可解释性。

3.3 RH 患者发生 MACE 的可解释性规则集的评价

为了验证本方法提取的规则集在准确性与可解释性方面的统计性能, 通过在 RH 测试集 611 例数据上的 3 个试验, 对决策树 (可以产生一组规则)、随机森林模型以及随机森林规则提取方法这三种模型的准确性、精度、灵敏度、特异度、F 值、受试者工作特征曲线下面积 (AUC 值) 以及可解释性进行评价, 结果如表 4 所示。结果显示, 三种模型的准确性、精度、灵敏度都较高, 但随机森林规则提取方法的可解释性 (0.96) 优于决策树的可解释性 (0.83), 且大大优于随机森林模型 (为负值), 表明借用了可解释树集成框架的规则提取方法获得的分类规则, 在准确性和解释性之间能够达到很好的平衡。

4 讨论

本研究结果显示, 存在“咳嗽”“老舌”“憋喘”“颈动脉硬化”“弦脉”等症状的 RH 患者容易发生 MACE。结合中医理论分析, “咳嗽”是因心阳不振导致血脉瘀阻、水饮停聚并累及肺脏, 当水饮射肺, 肺气上逆, 可致咳嗽; “老舌”多提示病邪较深, 侵入脏腑, 伤及阴血; “憋喘”提示心肺气虚, 正如《灵枢·五阅五使》曰: “故肺病者, 喘息鼻张”, 《灵枢·本脏》曰: “肺高则上气肩息咳”; “颈动脉硬化”“脉弦”则因痰、瘀、虚等导致血脉流滞不畅, 瘀阻日久而损伤血脉, 进而加重

表3 分类顽固性高血压合并主要不良心血管事件患者的简化规则集

规则	长度	频率	误差	条件	分类值
1	4	0.048	0	咳嗽c('1') & 老舌c('1') & 头晕c('0') & 胸闷c('0')	RH+MACE患者
2	5	0.018	0	憋气c('0') & 头晕c('1') & 胸闷c('1') & 苔腻c('0') & 年龄大于60岁c('0')	RH患者
3	3	0.050	0.027	恶心c('0') & 颈动脉硬化c('1') & 弦脉c('1')	RH+MACE患者
4	3	0.014	0.048	既往高血压c('0') & 憋喘c('0') & 痰量少c('1')	RH患者
5	6	0.024	0.083	精神欠佳c('0') & 咳嗽c('1') & 咯痰c('0') & 慢性肾功能不全c('0') & 胸闷c('0') & 苔腻c('0')	RH患者
6	1	0.121	0.045	憋喘c('1')	RH+MACE患者
7	3	0.167	0.169	既往糖尿病c('1') & 家族史高血压病c('0') & 腰痛c('0')	RH+MACE患者
8	5	0.105	0.154	既往脑梗死c('0') & 憋喘c('0') & 憋气c('0') & 弦脉c('1') & 年龄大于60岁c('0')	RH患者
9	2	0.070	0.194	高尿酸血症c('1') & 头痛c('0')	RH+MACE患者
10	5	0.197	0.226	既往脑梗死c('0') & 大便秘结c('0') & 反流性食管炎c('0') & 咳嗽c('0') & 慢性肾功能不全c('0')	RH患者
11	1	0.186	0.193	Else	RH+MACE患者

注：长度，影响因素的个数；频率，满足某一规则中规则条件的样本比例；误差，规则确定的错误结局事件的样本数除以满足规则条件的样本数；c('1')代表“有”此特征值；c('0')代表“无”此特征值；“&”代表特征值对“并”；Else，其他的规则条件；RH，顽固性高血压；MACE，主要不良心血管事件。

表4 顽固性高血压患者发生主要不良心血管事件的三种模型评价指标的比较

模型	准确性	精度	灵敏度	特异度	F值	AUC值	可解释性
决策树	0.85	0.89	0.85	0.85	0.86	0.84	0.83
随机森林	0.97	0.99	0.95	1.00	0.97	0.98	→0**
随机森林规则提取方法	0.81	0.83	0.85	0.73	0.84	0.80	0.96

注：AUC值，受试者工作特征曲线下面积；*表示可解释性为负值。

病理产物停留，弦脉又主气滞，《金匱要略·呕吐下利病脉证治》曰：“脉弦者虚也”，间接提示脉弦体征出现则病情发展较深，气血阴阳失调，气津两伤、血脉不畅较为明显。综上分析，RH发生MACE的病理特征与心肺两脏关联性大，病机总属虚实夹杂，因RH疾病日久，由实转虚，主责心肺气虚。心主血脉，肺朝百脉，心肺气虚则全身气血运行不畅，加重痰浊、血瘀等病理产物滞留经络，久蕴成毒，损伤心、肺、脑络脉，最终发生MACE。另外本研究还发现，RH患者无“头晕”“胸闷”“恶心”症状，则出现MACE的可能性更大。分析原因，可能为头晕、胸闷症状多是RH刚出现症状时的心脑疾病症状，伴随着患者用药干预及生活方式的改变等，后期头晕、胸闷等症状并不典型；“恶心”多因胃气上逆所致，痰饮易致脾胃气机阻滞，若不存在恶心的临床症状，证明脾胃症状不明显。

在考虑了中医四诊的疾病分类研究中，与传统

的解释性统计方法相比，随机森林、提升树、深度神经网络等集成学习算法被认为是优化预测或分类准确性的最实用的方法。但以上算法依然是“黑盒”模型，忽视了方法的不可解释性，且几乎没有应用于中医学研究，更没有关注中医四诊对疾病预测的含义^[21-23]。有研究认为应该停止将“黑盒”统计机器学习模型用于高风险决策，而应采用可解释性模型^[24]。但在统计机器学习领域，可解释性模型很难直接创建，通常都是创建第二个（事后）模型来解释第一个“黑盒”模型。规则提取方法通过对树集成模型的第二次“加工”（可解释树集成框架）提取出可解释性的规则，从而显示分类过程中可解释的中医四诊规则和模式，可帮助医生或研究者做出合理的且更好的“病证”结合治疗决策。因此，本研究基于EMR系统数据运用规则提取方法对RH+MACE的相关中医四诊信息进行可解释性研究。本研究中，规则提取方法相对于原始的“黑盒”模型，事后模型可解释性可能是原始模型在部分特征空间中的不准确表示，但本研究对规则提取方法的准确性和可解释性给予了度量，这在一定程度上缓解了统计机器学习“黑盒”模型在疾病预测或分类中的可解释性问题，也为进一步探索中医药诊治RH过程中“病-症-证-药”的精准化辨证思路及处方用药的关键技术提供了新的方法。

参考文献

- [1]WHELTON PK, CAREY RM, ARONOW WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/

- NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines[J]. J Amer Coll Cardiol, 2018, 71(19): e127-e248.
- [2] KASIAKOGIAS A, TSIOUFIS C, DIMITRIADIS K, et al. Dynamic resistant hypertension patterns as predictors of cardiovascular morbidity: a 4-year prospective study [J]. J Hypertens, 2014, 32(2): 415-422.
- [3] 刘大胜, 向兴华, 邢云飞, 等. 基于文本挖掘的顽固性高血压的治疗现状及“证-药-效”的规律研究[J]. 世界科学技术-中医药现代化, 2020, 22(3): 836-842.
- [4] YANG J, LI Y, LIU Q, et al. Brief introduction of medical database and data mining technology in big data era [J]. J Evid Based Med, 2020, 13(1): 57-69.
- [5] ZHANG J, KOWSARI K, HARRISON JH, et al. Patient 2Vec: a personalized interpretable deep representation of the longitudinal electronic health record[J]. IEEE Access, 2018, 6: 65333-65346. doi: 10.1109/ACCESS. 2018. 2875677.
- [6] STEELE AJ, DENAXAS SC, SHAH AD, et al. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease[J]. Plos One, 2018, 13(8): e0202344. doi: 10.1371/journal.pone.0202344.
- [7] 刘大胜. 真实世界下顽固性高血压病不良心血管事件的诊治规律研究[D]. 北京: 中国中医科学院, 2020.
- [8] FOKKEMA M. Fitting prediction rule ensembles with R package pre[J]. J Statist Software, 2020, 92(12): 1-30.
- [9] DENG H. Interpreting tree ensembles with in trees [J]. Int J Data Science Analyt, 2018, 7(4): 277-287.
- [10] HAILESILASSIE T. Rule extraction algorithm for deep neural networks: a review [J]. Int J Comput Science Inform Secur, 2016, 14(7): 376-381.
- [11] FRIEDMAN JH, POPESCU BE. Predictive learning via rule ensembles [J]. Ann Appl Stat, 2008, 2(3): 916-954.
- [12] 中国高血压防治指南修订委员会. 中国高血压防治指南 2010[J]. 中华心血管病杂志, 2011, 39(7): 579-616.
- [13] 《中国高血压防治指南》修订委员会. 中国高血压防治指南(2018年修订版)[J]. 中国心血管杂志, 2019, 24(1): 24-56.
- [14] 常宝宝. 自然语言分析与生成术语简介[J]. 术语标准化与信息技术, 2010(4): 19-22.
- [15] NGUYEN TT, NGUYEN MP, PHAM XC, et al. Heterogeneous classifier ensemble with fuzzy rule-based meta learner [J]. Informat Science, 2018, 422: 144-160. doi: 10.1016/j.ins.2017.09.009.
- [16] DENG H, RUNGER G, TUV E, et al. CBC: an associative classifier with a small number of rules [J]. Decis Support System, 2014, 59: 163-170. doi: 10.1016/j.dss.2013.11.004.
- [17] DENG H, RUNGER G. Gene selection with guided regularized random forest[J]. Pat Recognit, 2013, 46(12): 3483-3489.
- [18] FÜRNKRANZ J. Separate-and-conquer rule learning [J]. Artif Intell Review, 1999, 13: 3-54. doi: 10.1023/A:1006524209794.
- [19] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 347-352.
- [20] GORZAŁCZANY MB, RUDZIŃSKI F. A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability[J]. Appl Soft Comput, 2016, 40: 206-220. doi: 10.1016/j.asoc.2015.11.037.
- [21] 项莎特, 瞿溢谦, 叶含笑. 多层学习联合建模方法设计在气阴两虚型咳嗽证候的辨证诊断中的应用[J]. 中国卫生统计, 2020, 37(6): 892-894.
- [22] 丁亮, 章新友, 刘莉萍, 等. 基于深度神经网络的原发性肝癌证型诊断分类预测模型[J]. 世界科学技术-中医药现代化, 2020, 22(12): 4185-4192.
- [23] 崔伟锋, 刘萧萧, 韩静漪, 等. 基于随机森林的原发性高血压心血管风险预后模型[J]. 中国老年学杂志, 2020, 40(4): 814-816.
- [24] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nat Mach Intell, 2019, 5: 206-215. doi: 10.1038/s42256-019-0048-x.

Interpretability of Chinese Medicine Four Examinations Information of Major Adverse Cardiovascular Events in Resistant Hypertension: based on Random Forest Rule Extraction Method

XIANG Xinghua^{1,2}, PENG Yehui¹, YANG Wei², LIU Jian³, LIU Dasheng², JI Xinyu², LEI Shuyang², XIE Feibiao^{1,2}, YANG Jiwen², WANG Liying², HAN Xuejie², SHANG Hongcai⁴

1. School of Mathematics and Computational Science, Hunan University of Science and Technology, Xiangtan, 411201; 2. Institute of Basic researches in Clinical Medicine, China Academy of Chinese Medical Sciences; 3. Xiyuan Hospital, China Academy of Chinese Medical Sciences; 4. Dongzhimen Hospital, Beijing University of Chinese Medicine

ABSTRACT Objective To explore the classification rules and characteristics of clinical information in resistant

hypertension (RH) patients with major adverse cardiovascular events (MACE). **Methods** We collected the electronic medical records of 2503 RH patients, among which 1286 cases had MACE. The basic personal information including gender, age, past history, family history, personal history, and history of infectious diseases, traditional Chinese medicine (TCM) four examinations information, and western medicine diagnosis information were collected. The previous study identified 140 indicators closely related to the occurrence of MACE in RH patients. The distribution of the closely related indicators in 2093 patients was analyzed, and the random forest rule extraction method was used to analyze the characteristics that influence the occurrence of MACE in RH patients. Multiple decorrelation decision trees were built through structured data processing and random forest training, which can generate a wealth of decision rules. The rule extraction method was used to extract, measure, trim, and select rules. The sequential coverage method was used to filter the simplified classification rule set, and the interpretability measurement of the rules was defined. **Results** The top frequent five closely related indicators of the TCM four examinations information among 2093 patients were wiry pulse (1101, 52.6%), greasy coating (947, 45.25%), poor sleep (870, 41.57%), and pale complexion (784, 37.46%). The top five frequent information related to patients' basic information and western medicine diagnosis information were normal heart rate (1765, 84.33%), previous hypertension (1601, 76.49%), age older than 60 years old (1547, 73.91%), normal diastolic blood pressure (1428, 68.23%), and grade 3 hypertension (1151, 54.99%). The interpretability analysis of the influencing factors showed that were three rules for predicting RH+MACE with error less than 0.05, which were Rule I: cough c('1') & old tongue c('1') & dizziness c('0') & chest tightness c('0'); Rule II: Nausea c('0') & carotid arteriosclerosis c('1') & Xuanmai c('1'); and Rule III: suffocation c('1'). The classification accuracy of the three rules has reached more than 95% (error less than 0.05), and has a certain clinical interpretability. The model evaluation index results of the rule set: sensitivity (0.85), specificity (0.73), accuracy (0.83), area under the receiver operating characteristic curve (AUC) value (0.80), and interpretability value (0.96). **Conclusion** Presentation of cough, old tongue, carotid arteriosclerosis, wiry pulse, wheezing, and with no symptoms of dizziness, chest tightness and nausea are taken as the core classification elements for predicting MACE in RH patients.

Keywords resistant hypertension; major adverse cardiovascular events; random forest rule; interpretability studies

(收稿日期: 2021-03-15; 修回日期: 2021-11-15)

[编辑: 邓媛]

广告

欢迎邮购《中医杂志》过刊

2000、2001 年合订本每本 100 元, 2002、2003 年合订本每本 105 元, 2004 年合订本 115 元, 2005 年合订本 125 元, 2007—2010 年合订本每本 140 元。2011(缺 5、6 期)—2013 年每册 9.80 元, 2014(缺 21 期)—2017 年(缺 3、5、15、18 期)每册 15 元, 2018 年(缺 5、7、9 期)每册 20 元, 2020 年(缺 7、8、9 期)每册 20 元, 2021 年每册 20 元。2011—2013 年合订本每年 4 本, 全年 280 元, 2014—2017 年合订本每年 4 本(缺 2017 年上半年合订本), 全年 400 元, 2018—2022 年合订本, 每年 4 本, 全年 600 元。《中医杂志》2006—2008 年增刊每册 40 元。

以上免费邮寄, 如需挂号每件另加 3 元。

欲购者请汇款至北京市东直门内南小街 16 号, 收款人: 中医杂志社

或关注《中医杂志》官方微信(zyzgzgh)购买。邮编: 100700, 电话: (010)64089195。