

DOI:10.13288/j.11-2166/r.2023.24.008

结合关联规则的隐结构分析构建证候状态辨识模型 ——以原发性骨质疏松症肾阳虚证为例

栗贝贝^{1,2}, 王晶^{3,4}, 向兴华¹, 徐文源¹, 白卫国¹, 刘孟宇¹,
李玉坤¹, 吴晓坤², 王拥军^{3,4}✉, 杨伟¹✉

1. 中国中医科学院中医临床基础医学研究所, 北京市东城区东直门内南小街16号, 100700; 2. 华北电力大学数理系; 3. 上海中医药大学脊柱病研究所; 4. 上海中医药大学附属龙华医院

[摘要] **目的** 以原发性骨质疏松症 (POP) 肾阳虚证为例, 构建中医证候状态定量化辨识模型, 为证候诊断标准化研究提供方法学参考。**方法** 采用描述统计分析筛选出 POP 高频临床特征, 通过关联规则算法得到 POP 强关联特征。在此基础上结合隐结构分析法建立隐结构 (隐树) 模型, 对与 POP 肾阳虚证相关隐变量和显变量 (特征) 进行综合聚类, 通过互信息与累积信息覆盖度等指标对聚类结果进行诠释, 探究其主症、次症, 根据各隐类中特征出现概率推断 POP 肾阳虚证状态类别。依据状态类别计算临床特征分值和辨识阈值, 结合综合判断规则初步构建 POP 肾阳虚证状态辨识模型, 最后以中医专业人员判断结果为金标准进一步评价模型的辅助辨识效果。**结果** 得到 POP 强关联的 32 个特征, 进一步构建的隐树模型贝叶斯信息准则 (BIC) 评分为 -15291.93。根据互信息和累积信息覆盖度推出 POP 肾阳虚证的主症为骨软、神疲、舌淡、小便清、夜尿频多、肢冷、脉沉细、苔白, 次症为乏力、性欲减退、大便溏、尿频、腰膝酸软、畏寒。由 POP 肾阳虚证不同隐类中各临床特征出现的概率, 推出 POP 肾阳虚证状态为 S0 类 (无/轻度肾阳虚证) /S1 类 (中度肾阳虚证) /S2 类 (重度肾阳虚证)。优化状态辨识初步规则, 细化状态 S1 类, 结果显示 970 例 POP 患者中, 520 例患者为无/轻度肾阳虚证、224 例为中度偏轻度肾阳虚证、81 例为中度偏重度肾阳虚证、145 例为重度肾阳虚证。模型评估指标正确率为 0.8835、灵敏度为 0.7181、特异度为 0.9437。**结论** 结合关联规则的隐结构分析可构建 POP 肾阳虚证状态辨识模型, 且该模型具有良好的定量化辨识效果, 可辅助临床医生提高中医诊断效率和准确性。

[关键词] 原发性骨质疏松症; 关联规则; 隐结构 (隐树); 肾阳虚证; 状态辨识模型

随着现代科学技术的发展以及数据驱动方法的广泛应用, 中医学领域出现了许多定量的证候标准研究, 如通过有监督统计机器学习方法构建证候辨识模型^[1], 或利用无监督方法 (比如聚类分析、隐结构分析) 进行辨证^[2]。但此类研究还存在两个方面的不足, 一是对中医证候的多个症状指标之间的关联性考虑不充分, 纳入了大量的中医特征, 使得模型分析过于复杂且可能影响辨证效果; 二是在证

候状态辨识方面不够深入, 只停留在辨证分型而忽视了状态辨识。关联规则和隐结构分析是常见的数据挖掘方法, 其中关联规则常用于获取症状变量之间的关联性^[3], 隐结构分析可以根据患者症状数据进行特征聚类分析, 以确定证候类型和状态分布^[4]。为了充分考虑多个特征之间的关联性, 减少非必要特征的纳入, 并获取证候的状态类别, 本研究结合了这两种方法的特点, 在一定程度上可降低模型复杂度和提高模型准确性。

在人口老龄化背景下, 骨质疏松已成为社会公共健康问题^[5]。原发性骨质疏松症 (primary osteoporosis, POP) 是由骨吸收大于骨形成导致的骨重

基金项目: 国家重点研发计划 (2018YFC1704306, 2018YFC1704300); 中国中医科学院科技创新工程项目 (CI2021A04706, CI2021B003); 中国中医科学院自主选题项目 (Z0643, Z0723)

✉ 通讯作者: 王拥军, yjwang8888@126.com; 杨伟, yangyxq@ruc.edu.cn

建失衡，其特点是骨小梁微结构破坏、骨强度下降及骨折风险增加^[6]。POP 临床一般按照肾阳虚证、肾阴虚证、脾肾阳虚证、肝肾阴虚证、肾阴阳两虚证等证型进行论治^[7-8]，研究调查同时也证实肾阳虚证或脾肾阳虚证为 POP 主要证型^[9]。因此，本研究基于 POP 患者的横断面调查数据，采用关联规则找到 POP 强关联的特征，联合隐结构分析法构建 POP 肾阳虚证状态辨识模型，为证候诊断标准化提供方法学支撑。本研究遵循伦理学原则，保护患者个人隐私。

1 资料与方法

1.1 诊断标准

POP 诊断标准参照《中国老年骨质疏松诊疗指南（2018）》^[10]和《中国定量 CT（QCT）骨质疏松症诊断指南（2018）》^[11]制定：双能 X 线测量股骨近端及腰椎骨密度，结果与健康成年人峰值骨量相比下降 2.5 个或更多标准差即可诊断。

1.2 纳入标准

符合上述诊断标准；年龄≥25 岁；意识清晰且具备阅读能力，与调查人员沟通无障碍；了解调查目的后签署知情同意书。

1.3 排除标准

同时患有其他疾病可影响中医证候判断的患者；有精神障碍、认知障碍者；临床症状、体征和舌脉变量取值有缺失的患者。

1.4 证型判定标准

本研究由 3 位工作年限≥5 年且具有高级职称的中医专业人员同时对每位患者进行中医证型判定。参考《中医诊断学》^[12]129-130 制定肾阳虚证辨证标准：腰膝酸软冷痛、畏寒肢冷，兼有神疲乏力、小便清、大便溏薄、夜尿频。以至少两位中医专业人员判定相同的结果作为人为判定结果，人为判定后再进行模型判定。如人为判定结果与模型判定结果不同，请中医专业人员再次进行判定，并确定最终人为判定结果。以最终人为判定结果与模型判定结果进行辨识模型的正确率、灵敏度等的验证。

1.5 研究对象

本研究共纳入 POP 患者 970 例，分别来自于 2010 年 8 月至 2013 年 10 月上海中医药大学附属龙华医院（679 例）、上海中医药大学脊柱病研究所（291 例）。一般情况详见表 1。

1.6 资料收集与分析

采用骨质疏松症临床症状调查问卷^[13]搜集研

表 1 纳入研究的 970 例原发性骨质疏松症患者一般情况

Table 1 General information of the included 970 patients with

primary osteoporosis	
项目	例数（%）
性别	
男	104（10.7）
女	865（89.3）
年龄	
<60 岁	277（28.7）
≥60 岁且<70 岁	347（36.0）
≥70 岁	340（35.3）
教育水平	
小学及以下	191（19.9）
中学（含中专）	569（59.2）
大学（含大专）及以上	201（20.9）
中医专业人员辨证	
肾阳虚证	259（26.7）
非肾阳虚证	711（73.3）

注：“性别”取值缺失 1 例；“年龄”取值缺失 6 例；“教育水平”取值缺失 9 例。

究对象的相关信息。在严格的质量控制下，对临床症状、体征和舌脉等 51 个特征以及一般资料进行采集与分析。

为了保证数据质量，在调查开始前对调查员进行调查技巧、医学知识等方面的统一培训，规范调查程序。调查过程中对被调查者不理解的条目作必要的解释工作，调查表的内容以被调查者自填和调查员的客观观察为准，不随意更改。调查表填写应认真、完整，数据如确有更改必须详细注明更改原因。

1.7 数据分析与统计方法

采用统计软件 R 4.2.0 中的 arules 包和隐结构分析软件 Lantern 5.0 进行统计分析。首先利用描述分析筛选出发生频率≥10% 的临床症状、体征和舌脉等特征（变量）。然后在关联规则 Apriori 算法中设定最小支持度 7.5%、最小置信度 80%、提升度>1 和项集项数≥2，可视化满足条件的有效强关联规则，尽可能多地保留 POP 强关联的特征^[14]。采用双步隐树分析（LTM-EAST）算法对强关联特征进行隐结构分析，以贝叶斯信息准则（BIC）评分（以负分计量）来评价构建的隐树模型，分值越大提示模型拟合越好^[15]。对反映肾阳虚证不同侧面的隐变量和显变量进一步综合聚类，根据互信息和累积信息覆盖度指标诠释 POP 肾阳虚证的主症和次症，参考各隐类中特征出现概率，推测肾阳虚证状态类别，并根据部分特征的分值、状态辨识阈

值和综合判断规则构建 POP 肾阳虚证状态辨识模型。

1.7.1 关联规则 关联规则中的三个重要指标为支持度、置信度和提升度,前两者反映规则的频繁性,后者反映规则的相关性^[16]。关联规则可用于实现对 POP 患者特征的关联,并在规则集中找到符合最小支持度、最小置信度、提升度>1 的强关联规则,进一步得到 POP 强关联的特征,其表达式为“A=>B”,例如 {小便清,肢冷}=> {舌淡} 表明了三者之间的强关联性,即出现 {小便清,肢冷} 的症状规则集中同时伴有 {舌淡} 的概率较高。本文采用关联规则 Apriori 算法基于最小支持度 7.5%、最小置信度 80%、提升度>1 和项集项数≥2 找出这些有意义的强关联规则,将其可视化得到 POP 强关联的特征。

1.7.2 隐结构分析 隐结构分析本质上属于隐类分析,当变量属性为分类变量时,它是一种基于模型的多维同时聚类方法,主要包括一系列隐结构(隐树)模型:隐类模型、分层隐类模型、多层隐结构模型、一般隐结构模型^[17]。在隐结构分析中,多个隐变量可能反映同一个证候的不同侧面^[18],可进一步对多个隐变量进行综合聚类,并利用互信息、信息覆盖度、累积互信息、累积信息覆盖度、隐类概率等统计指标对综合聚类产生的新隐变量进行诠释^[19],以判断 POP 肾阳虚证的核心症状,进一步用于状态辨识研究。采用隐结构分析法对 POP 强关联特征构建模型并对反映肾阳虚证不同侧面的隐变量和显变量综合聚类,对聚类结果诠释从而推测肾阳虚证状态类别,最终建立肾阳虚证状态辨识模型。

1.7.3 临床特征分值和辨识阈值 状态辨识模型是基于朴素贝叶斯原理确立临床特征分值和辨识阈值来构建。假设 Z 是隐变量,取值为 0 或 1, $X_1, X_2, \dots, X_m$ 为 m 个特征(变量),各特征的取值为 0 或 1, $\mathbf{X} = (X_1, X_2, \dots, X_m)$ 。 X_i 和 Z 之间的互信息 I 可以表示为

$$I(Z; X_i) = H(Z) - H(Z|X_i) \quad [20]$$

$$H(Z) = -\sum_z P(Z) \log P(Z) \quad [20]$$

$$H(Z|X_i) = -\sum_{x_i, z} P(X_i, Z) \log P(Z|X_i) \quad [20]$$

其中 $P(X_i, Z)$ 为 X_i 和 Z 的联合概率分布, $P(X_i)$ 和 $P(Z)$ 分别为 X_i 和 Z 的概率分布, $P(Z|X_i)$ 是给定 X_i 的 Z 的条件概率分布, $H(Z)$ 是 Z 的熵, $H(Z|X_i)$ 是给定 X_i 的 Z 的条件熵。

各临床特征分值表示为

$$\text{score}(X_i) = \log \frac{P(X_i|Z=1)}{P(X_i|Z=0)}, i = 1, 2, \dots, m \quad [21]$$

辨识患者状态的阈值表示为

$$\text{threshold}(Z) = \log \frac{P(Z=0)}{P(Z=1)} \quad [21]$$

2 结果

2.1 临床特征频次/频率分析

对 970 例 POP 的 51 个临床症状、体征和舌脉信息进行频次/频率统计,结果显示,频率≥10% 的特征(症状、体征和舌脉)共有 43 个,见表 2。

2.2 关联规则分析

图 1 示,通过关联规则分析,得到 POP 患者强关联的 32 个特征,包括性欲减退、烘热、面赤、脉细数、少苔、失眠、舌红、咽干、心悸、心烦、大便干结、小便短赤、腰膝酸软、神疲、骨软、齿脱、头发白、夜尿频多、尿频、多梦、苔白、气短、健忘、乏力、畏寒、肢冷、大便溏、纳差、反应迟钝、舌淡、脉沉细、小便清。

2.3 隐结构分析

2.3.1 隐结构(隐树)模型学习及评价 利用 LTM-EAST 算法对强关联的 32 个特征学习得到隐结构模型,其 BIC 评分为-15 291.93,优于 43 个特征的全模型(BIC 评分为-20 852.83)。图 2 示,隐结构模型中共引入 20 个隐变量,分别为 Y0-Y19。各个隐变量括号内的取值表示隐类(即隐变量状态)个数, Y2 包含 3 个隐类,其余 19 个隐变量均包含 2 个隐类。失眠、心烦等具体的临床特征为显变量。

2.3.2 综合聚类 由研究者根据《中医诊断学》^[12]173-176 结合自身经验判断,认为隐变量 Y9、Y11、Y14、Y18、Y19,显变量“舌淡”“苔白”“脉沉细”“性欲减退”都与中医证候肾阳虚证有关,从不同角度诠释了肾阳虚证的表现。通过对隐变量和显变量进行综合聚类,得到新的隐变量结果,详见图 3。可见隐变量 Z1 将 POP 人群划分为 3 类, Z1 可以解释为“肾阳虚证”隐变量。

2.3.3 POP“肾阳虚证”隐变量(Z1)诠释 Z1 中各临床特征对应的互信息值、累积互信息值、信息覆盖度以及累积信息覆盖度等如图 4 和表 3 所示。图 4 的横坐标中“骨软”排在最前,提示“骨软”与 Z1 的互信息(0.34)最高,即说明 Z1 的隐类在“骨软”特征上差别最大,其余临床特征与 Z1 的互信息见表 3。从累积信息覆盖度来看,到苔白时,累积信息覆盖度已经达到 95%,说明前 8 个特征

表 2 970 例原发性骨质疏松症患者常见临床特征
(≥10%) 分布

Table 2 Distribution of common clinical features (≥10%)
in 970 patients of primary osteoporosis yang deficiency

特征 (变量)		频次 (%)
症状体征	头发白	773 (79.7)
	健忘	670 (69.1)
	乏力	658 (67.8)
	齿脱	652 (67.2)
	腰膝酸软	627 (64.6)
	失眠	522 (53.8)
	畏寒	516 (53.2)
	多梦	505 (52.1)
	神疲	503 (51.9)
	口干	489 (50.4)
	面色少华	400 (41.2)
	咽干	373 (38.5)
	肢冷	365 (37.6)
	骨软	340 (35.1)
	性欲减退	336 (34.6)
	反应迟钝	324 (33.4)
	耳鸣	322 (33.2)
	眩晕	312 (32.2)
	手足心热	288 (29.7)
	大便干结	278 (28.7)
	心烦	276 (28.5)
	耳聋	263 (27.1)
	夜尿频多	263 (27.1)
	纳差	258 (26.6)
	面色萎黄	223 (23.0)
	气短	216 (22.3)
	尿频	208 (21.4)
	盗汗	196 (20.2)
	潮热	180 (18.6)
	面赤	178 (18.4)
	心悸	159 (16.4)
	烘热	150 (15.5)
	大便溏	147 (15.2)
	小便短赤	140 (14.4)
	小便清	128 (13.2)
	口唇淡白	111 (11.4)
舌脉	舌红	542 (55.9)
	脉沉细	429 (44.2)
	苔薄白	386 (39.8)
	脉细数	292 (30.1)
	少苔	243 (25.1)
	舌淡	233 (24.0)
	苔白	184 (19.0)

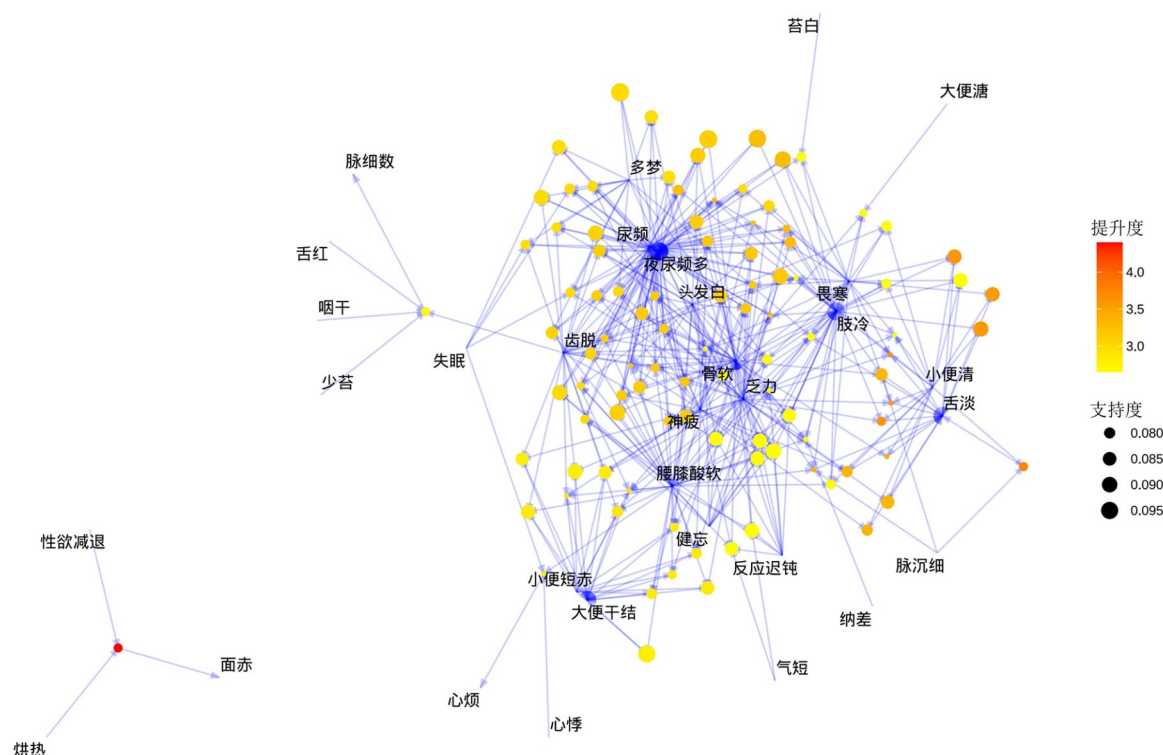
(骨软、神疲、舌淡、小便清、夜尿频多、肢冷、脉沉细、苔白) 可以作为肾阳虚主症, 后 6 个特征(乏力、性欲减退、大便溏、尿频、腰膝酸软、畏寒) 可作为次症。

2.3.4 POP “肾阳虚证” 隐变量 (Z1) 状态 使用隐结构分析继续对 Z1 进行模型诠释。如图 5 与表 4 所示, Z1 的三个隐类分别为 S0、S1 和 S2。S0 类患者在 POP 人群中的发生概率为 53%, 这类患者脉沉细出现概率稍高, 其他 13 个特征出现概率均相对较低, 可以认为 S0 类状态患者具有 “无/轻度肾阳虚证” 表现; S1 类患者的发生概率为 33%, S1 类患者中部分特征 (即骨软、神疲、乏力、性欲减退、畏寒、腰膝酸软) 出现的概率较高, 可以认为 S1 类状态患者具有 “中度肾阳虚证” 表现; S2 类患者的发生概率为 14%, 此类患者中除尿频外, 其他特征出现概率均相对较高, 可以认为 S2 类患者具有 “重度肾阳虚证” 表现。

2.3.5 构建 POP 肾阳虚证状态辨识模型 通过对 Z1 的状态分析, 得到其状态的临床症状、舌脉分值和肾阳虚证状态辨识阈值, 根据症状和舌脉分值总和与阈值大小的比较规则可以判断患者的状态类别。表 5 中的 “分值 1” 是在判断某一个患者属于 S1 还是 S0 类人群时的各特征分值, 此处的辨识阈值为 10, 即当某个患者的症状和舌脉总分值大于 10 时, 则将其判别为 S1 类, 反之则为 S0 类; “分值 2” 是在判断某一个患者属于 S2 还是 S0 类人群时的各症状和舌脉分值, 即当某个患者的症状和舌脉总分值大于 31 时, 则将其判别为 S2 类, 反之则为 S0 类; “分值 3” 是在判断某一个患者属于 S2 还是 S1 类人群时的特征分值, 当某个患者的症状和舌脉总分值大于 19 时, 将其判别为 S2 类, 反之则为 S1 类。

根据表 5 中肾阳虚证的症状和舌脉分值以及辨识阈值, 首先判断每个患者是属于 S1 还是 S0 类, 然后判断每个患者是属于 S2 类还是 S0 类, 接着判断每个患者是属于 S2 还是 S1 类。对每位患者都有 3 次判断, 根据状态辨识的初步规则, 结果显示, 状态为 S1 类的 POP 患者中有 84 例为肾阳虚证, 221 例为非肾阳虚证, 即 S1 类中肾阳虚证和非肾阳虚证的例数比约为 1 : 3, 因此有必要对状态 S1 进行更为精准的划分, 将其划分为中度偏轻度肾阳虚证 (S1-0 类) 和中度偏重度肾阳虚证 (S1-2 类)。状态辨识的初步规则和优化后规则见表 6。

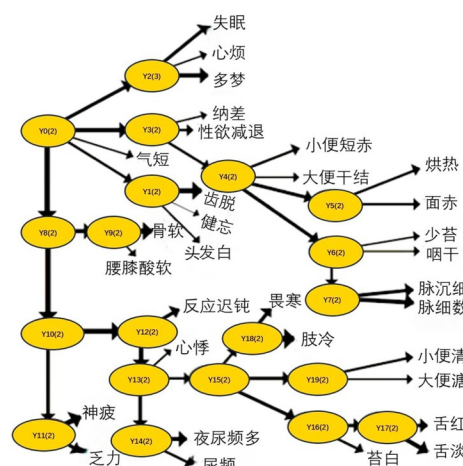
利用表 6 的优化后规则对 POP 患者再次进行状态辨识, 结果见表 7。在 970 例 POP 患者中, 520



注：图中箭头的末端（非尖角状）表示规则左项，蓝色箭头表示关系指向的方向，箭头指向的首端（尖角状）表示该规则的后继，即特征间的支持度、提升度大小或规则右项。圆圈颜色代表此规则提升度，由浅黄色到红色，提升度不断增加；圆圈大小代表此规则支持度，圆圈越大，支持度越大；圆圈颜色代表此规则提升度，颜色越深，提升度越高。置信度均在80%及以上。

图1 970例原发性骨质疏松症患者临床症状、体征和舌脉的关联规则可视化结果

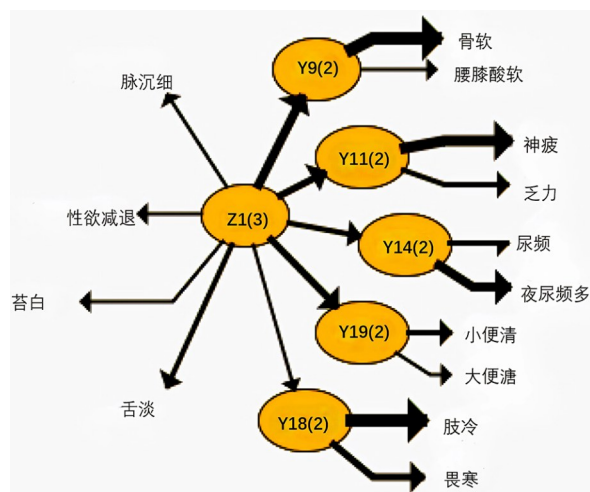
Figure 1 Visualisation results of association rules for clinical symptoms, signs, tongue and pulse in 970 patients with primary osteoporosis



注：每个隐变量代表对数据的一种划分，例如Y7(2)表示隐变量Y7有两个取值，代表将原发性骨质疏松症患者样本划分为两个隐类（隐变量状态）。图中的连线称为“边”，每一条边所连接的两个节点是直接相关的，每条边的粗细显示两个节点之间关联度的高低，边越粗，关联度越高。

图2 970例原发性骨质疏松症患者32个强关联特征的隐树模型

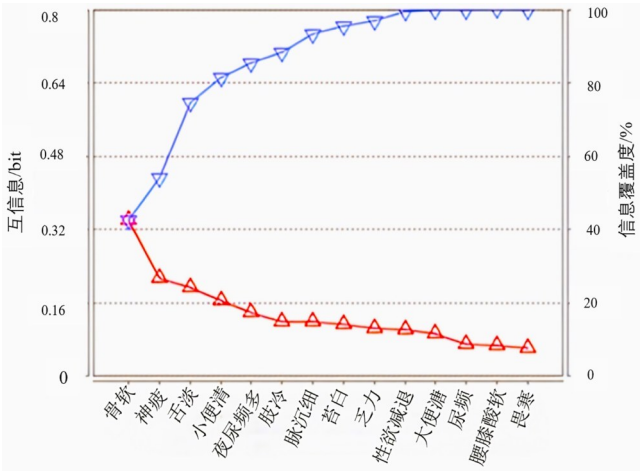
Figure 2 Latent tree model for 32 strongly correlated features in 970 patients with primary osteoporosis



注：Z1是隐变量与显变量综合聚类后得到的“肾虚证”隐变量；箭头的粗细显示两个变量（隐变量和隐变量或隐变量和显变量）之间关联度的高低，箭头越粗，关联度越高。

图3 原发性骨质疏松症患者部分隐变量与显变量的综合聚类图

Figure 3 Pooled clustering of selected implicit and explicit variables in patients with primary osteoporosis



注：横坐标是“肾阳虚证”隐变量（Z1）对应的临床特征；红色曲线为互信息（以左纵坐标为准）或信息覆盖度（以右纵坐标为准），蓝色曲线为累积互信息（以左纵坐标为准）或累积信息覆盖度（以右纵坐标为准）。

图4 原发性骨质疏松症患者“肾阳虚证”隐变量（Z1）的信息曲线图

Figure 4 Information curve of the implicit variable (Z1) for “kidney yang deficiency” in primary osteoporosis patients

表3 原发性骨质疏松症患者“肾阳虚证”隐变量（Z1）的互信息及信息覆盖度

Table 3 Mutual information and information coverage of the implicit variable (Z1) for “kidney yang deficiency” in primary osteoporosis patients				
肾阳虚证 隐变量(Z1)	互信息/bit	累积互信 息/bit	信息覆盖 度/%	累积信息覆 盖度/%
骨软	0.34	0.34	43	43
神疲	0.21	0.43	27	54
舌淡	0.19	0.60	24	75
小便清	0.16	0.65	21	81
夜尿频多	0.14	0.68	18	86
肢冷	0.12	0.71	15	88
脉沉细	0.12	0.75	15	93
苔白	0.11	0.76	14	95
乏力	0.11	0.77	13	97
性欲减退	0.10	0.79	13	99
大便溏	0.09	0.80	12	100
尿频	0.07	0.80	9	100
腰膝酸软	0.07	0.80	8	100
畏寒	0.06	0.80	8	100

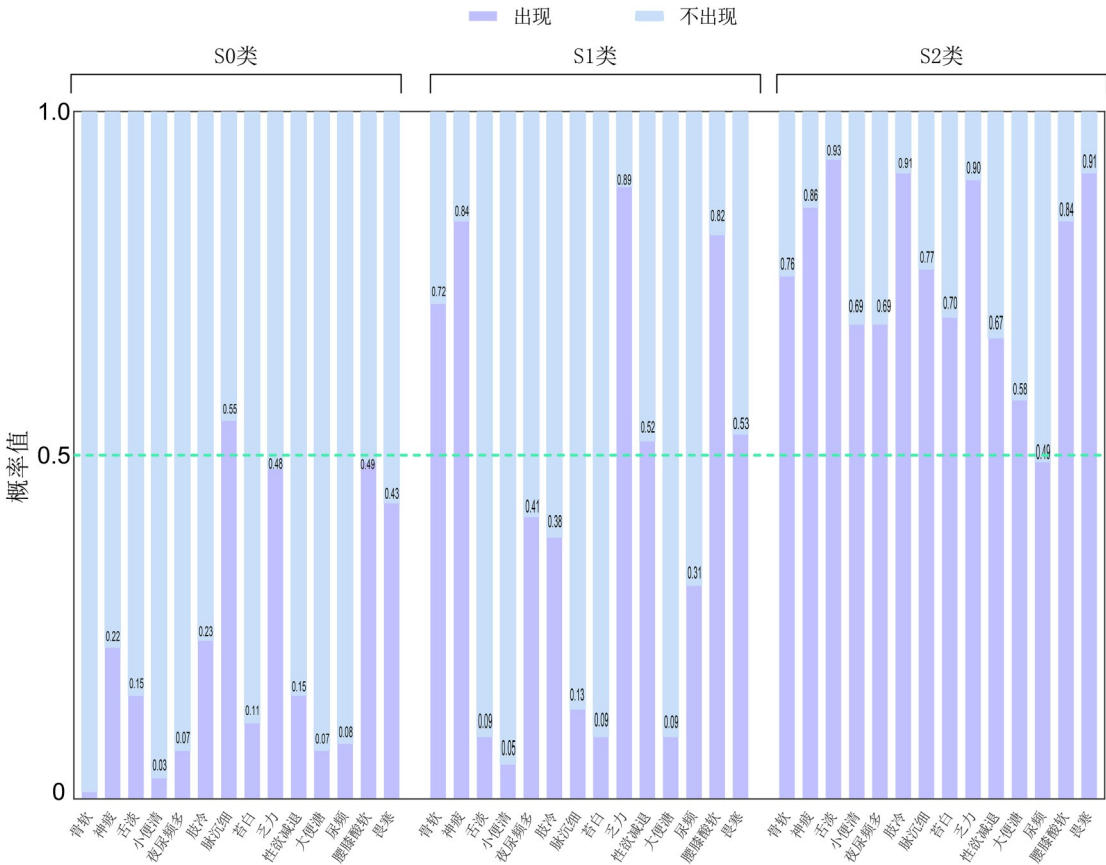


图5 原发性骨质疏松症患者“肾阳虚证”隐变量（Z1）的S0/S1/S2类人群中临床特征的条件概率

Figure 5 Conditional probabilities of clinical characteristics in the S0/S1/S2 categories for the implicit variable (Z1) for “kidney yang deficiency” in primary osteoporosis patients

表 4 原发性骨质疏松症患者“肾阳虚证”隐变量 (Z1) 各隐类概率及临床特征的条件概率

Table 4 Probability of each latent category and conditional probability of clinical characteristics of the implicit variable (Z1) for “kidney yang deficiency” in primary osteoporosis patients

肾阳虚证 (Z1) 症状体征	P=0.53 (Z1=S0类)		P=0.33 (Z1=S1类)		P=0.14 (Z1=S2类)	
	不出现概率	出现概率	不出现概率	出现概率	不出现概率	出现概率
骨软	0.99	0.01	0.28	0.72	0.24	0.76
神疲	0.78	0.22	0.16	0.84	0.14	0.86
舌淡	0.85	0.15	0.91	0.09	0.07	0.93
小便清	0.97	0.03	0.95	0.05	0.31	0.69
夜尿频多	0.93	0.07	0.59	0.41	0.31	0.69
肢冷	0.77	0.23	0.62	0.38	0.09	0.91
脉沉细	0.45	0.55	0.87	0.13	0.23	0.77
苔白	0.89	0.11	0.91	0.09	0.30	0.70
乏力	0.52	0.48	0.11	0.89	0.10	0.90
性欲减退	0.85	0.15	0.48	0.52	0.33	0.67
大便溏	0.93	0.07	0.91	0.09	0.42	0.58
尿频	0.92	0.08	0.69	0.31	0.51	0.49
腰膝酸软	0.51	0.49	0.18	0.82	0.16	0.84
畏寒	0.57	0.43	0.47	0.53	0.09	0.91

表 5 原发性骨质疏松症患者“肾阳虚证”隐变量 (Z1) 的临床特征分值及辨识阈值举例

Table 5 Examples of clinical characteristic scores and identification thresholds for the implicit variable (Z1) for “kidney yang deficiency” in primary osteoporosis patients

症状体征	分值1	分值2	分值3
骨软	7.9	8.0	0.0
神疲	4.2	5.0	0.0
乏力	3.1	3.0	0.0
脉沉细	-2.9	2.0	5.0
夜尿频多	3.3	5.0	2.0
性欲减退	2.6	3.0	1.0
腰膝酸软	2.3	3.0	0.0
尿频	2.4	4.0	1.0
肢冷	1.0	5.0	4.0
畏寒	0.6	4.0	3.0
舌淡	-0.7	7.0	7.0
苔白	-0.4	5.0	5.0
小便清	0.6	6.0	6.0
大便溏	0.2	4.0	4.0
辨识阈值	10	31	19

表 6 原发性骨质疏松症患者肾阳虚证状态辨识的初步规则和优化后规则

Table 6 Initial and optimised rules for identifying the state of kidney yang deficiency in patients with primary osteoporosis

患者属于		患者属于		患者属于		状态辨识初步规则	状态辨识优化后规则
S1类	S0类	S2类	S0类	S2类	S1类		
√		√		√		S2	S2
√		√			√	S1	S1-2
√			√	√		进一步判断	进一步判断
√			√		√	S1	S1-0
	√	√		√		S2	S2
	√	√			√	进一步判断	进一步判断
	√		√	√		S2	S2
	√		√		√	S0	S0

注：判断规则举例说明，若某位患者在3次辨识中分别为S1类、S2类、S2类，则根据优化后规则，将其状态辨识为S2类；若某位患者在3次辨识中分别为S1类、S0类、S1类，则根据优化后规则，将其状态辨识为S1-0类。

表 7 970例原发性骨质疏松症患者辨证结果统计

Table 7 Statistics on the results of syndrome differentiation for 970 patients with primary osteoporosis [例(%)]

患者	例 (%)	辨识为S0	辨识为S1-0	辨识为S1-2	辨识为S2
肾阳虚证	259 (26.70)	45 (8.65)	28 (12.50)	56 (69.14)	130 (89.66)
非肾阳虚证	711 (93.30)	475 (91.35)	196 (87.50)	25 (30.86)	15 (10.34)
合计	970 (100)	520 (100)	224 (100)	81 (100)	145 (100)

例属于无/轻度肾阳虚证 (S0), 224 例属于中度偏轻度肾阳虚证 (S1-0)、81 例属于中度偏重度肾阳虚证 (S1-2), 145 例属于重度肾阳虚证 (S2)。S0 类患者中 91.35% 为非肾阳虚证, S1-0 类患者中 87.50% 为非肾阳虚证, S1-2 类患者中 69.14% 为肾阳虚证, S2 类患者中 89.66% 为肾阳虚证, 这说明状态辨识的优化后规则更为合理与精准。

为进一步评价 POP 肾阳虚证状态辨识模型效果, 将 S1-2 类和 S2 类视为“肾阳虚证”状态, S0 类和 S1-0 类视为“非肾阳虚证”状态, 以中医专业人员判断结果为金标准, 则其与模型都判断为“肾阳虚证”状态的患者有 186 例、都判断为“非肾阳虚证”状态的患者有 671 例, 两者判断结果一致的患者共 857 例, 故优化后模型正确率为 0.8835、灵敏度为 0.7181、特异度为 0.9437, 详见表 8。

表 8 中医专业人员与辨识模型对 970 例原发性骨质疏松症患者肾阳虚辨证结果比较
Table 8 Comparison of the results of syndrome differentiation for 970 patients with primary osteoporosis by TCM professionals and the model [例 (%)]

患者	例数	辨识为“肾阳虚证”状态	辨识为“非肾阳虚证”状态
肾阳虚证	259	186 (71.81)	73 (28.19)
非肾阳虚证	711	40 (5.63)	671 (94.37)

3 讨论

临床中 POP 证型繁多且常相兼为病, 医师辨证技能水平很大程度取决于经验, 通过数据挖掘为 POP 肾阳虚状态建立客观定量标准的研究方法可以有效地辅助医生辨证论治。关联规则与隐结构分析已分别被广泛应用于中医临床数据研究^[22-23], 但将两者结合运用进行证候状态辨识仍不多见。关联规则可揭示临床特征间的关联关系并得到核心特征, 但难以判断患者的证候状态。隐结构模型可以客观量化地揭示临床特征与证候间的复杂关系, 但是特征变量数量过多会影响模型构建效率。因此, 本研究将关联规则与隐结构分析联合运用以弥补单一方法的不足。本研究在建立隐结构模型前, 采用描述统计和关联分析的方法对临床特征进行筛选, 获得出现频率高且强关联的 POP 特征变量, 而既往研究中普遍多以频次/频数分析来筛选将纳入隐结构模型的临床特征^[24-26]。本研究采用的隐结构分析法能够全面揭示 POP 临床特征与证候状态间的关系, 相比于隐类分析放宽了局部独立假设, 可以更

好地拟合数据, 更贴近中医理论^[15]。

在构建 POP 肾阳虚状态辨识模型过程中, 我们发现与 POP 强关联的特征包括腰膝酸软、骨软、齿脱、头发白等, 符合现代中医学“肾虚证”的特征。肾主骨, 生髓, 《金匱翼·腰痛》云:“肾虚腰痛者……其症形羸气少, 行立不支, 而卧息少可。无甚大痛, 而悠悠戚戚, 屡发不已”, 可印证 POP 肾虚导致髓化无源, 骨骼不能充养^[27], 即“肾虚髓枯”的中医病机^[28], 提示将聚类分析的方法引入中医学证候研究是可行的。

本研究在使用隐结构与综合聚类分析得出比较契合临床实际的聚类结果的基础上, 进一步尝试了利用辨证规则对 POP 肾阳虚证的不同状态进行辨识并结合中医专业人员辨证结果, 发现无/轻度、中度偏轻度患者中绝大多数为“非肾阳虚证”状态, 体现了该辨识模型在辨识非肾阳虚证方面的良好性能; 重度患者中绝大部分是“肾阳虚证”状态, 同样体现了此模型在该状态辨识中的准确性, 临床上有可能为医生的初步辨病辨证提供客观支持。但是在中度偏重度状态的患者中, 此模型的辨识效果不如其他三种状态, 肾阳虚证例数比例并未明显大于非肾阳虚证。其原因可能有二: 一是随着年龄的增加, 肾气衰, 天癸竭, 更易出现各脏皆弱、肾阴阳两虚等, 而非单见肾阳虚一证, 导致本研究中判别为单纯肾阳虚证例数减少; 二是病历在形成时中医专业人员与患者两方均存在一定的主观因素, 中医专业人员辨证一定程度上依据主观经验, 可能对同一组症状相对应的证候有不同的认识, 患者首诊时对于某些症状的描述也可能存在不太准确的主观描述。

本研究采用关联规则和隐结构分析相结合的数据挖掘方法构建 POP 肾阳虚证状态辨识模型, 与未考虑关联分析的隐结构模型相比, 拟合度与复杂度均较好, 可运用于个体 POP 肾阳虚证状态类别推测, 进一步明确临床特征与证候状态之间的关系, 也可为中医证候量化研究提供新思路。需要指出的是, 本研究中关联规则中支持度和置信度最小值的设定, 是以模型复杂度和合理性为出发点, 其客观依据有待进一步探讨。此外, 本研究中临床特征的取值均为二分类而非多分类, 无法量化症状的轻重程度, 故而模型的预测精度仍有待今后的研究继续提高。

参考文献

- [1] 陈星, 陈璟, 陈超, 等. 基于气相色谱-声表面波传感器联用技术的中医脾胃证候辨识[J]. 世界中医药, 2020, 15(11): 1540-1545.
- [2] 白琛, 冉春龙, 李梦君, 等. 基于隐结构结合因子分析探讨豫东农村地区脑梗死患者的常见证候[J]. 中国实验方剂学杂志, 2022, 28(7): 190-197.
- [3] 闫朝升, 朱彦峰, 白晶, 等. 数据挖掘及其在中医临床疗效评价中的应用研究[J]. 中医药学报, 2012, 40(3): 185-187.
- [4] 张连文, 袁世宏, 陈翌, 等. 隐结构模型与中医辨证研究(Ⅱ): 肾虚数据分析[J]. 北京中医药大学学报, 2008, 31(9): 4.
- [5] 杨惠林. 健康老龄化十年, 聚焦骨质疏松症[J]. 中华骨与关节外科杂志, 2022, 15(9): 652-655.
- [6] 中华医学会骨质疏松和骨矿盐疾病分会. 原发性骨质疏松症诊疗指南(2017)[J]. 中国骨质疏松杂志, 2019, 25(3): 281-309.
- [7] 林贤灿, 陈桐堂, 林燕平, 等. 原发性骨质疏松症的中医认识与探索[J]. 中国骨质疏松杂志, 2023, 29(6): 870-874, 885.
- [8] 温经渊, 曾晗冰, 吴连国. 骨质疏松症中医辨证分型研究进展[J]. 安徽中医药大学学报, 2021, 40(5): 101-104.
- [9] 曾昭洋, 胡文斌, 魏学玲, 等. 中老年人群原发性骨质疏松中医体质及辨证分型分布[J]. 中国老年学杂志, 2018, 38(2): 435-438.
- [10] 《中国老年骨质疏松诊疗指南》(2018)工作组, 中国老年学和老年医学学会骨质疏松分会, 马远征, 等. 中国老年骨质疏松诊疗指南(2018)[J]. 中国老年学杂志, 2019, 39(11): 2557-2575.
- [11] 《中国定量 CT(QCT) 骨质疏松症诊断指南》工作组, 程晓光, 王亮, 等. 中国定量 CT(QCT) 骨质疏松症诊断指南(2018)[J]. 中国骨质疏松杂志, 2019, 25(6): 733-737.
- [12] 李灿东, 黄惠勇, 王忆勤, 等. 中医诊断学[M]. 4版. 北京: 中国中医药出版社, 2016.
- [13] 邓洋洋, 王梅, 吕爱平, 等. “生长壮老取决于肾”流行病学调查信息采集表的信度和效度评价[J]. 中华中医药杂志, 2014, 29(6): 1944-1948.
- [14] 杨柳, 徐晴文, 黄汝佳, 等. 关联规则在中医药数据挖掘中的应用[J]. 中医药信息, 2022, 39(12): 35-40.
- [15] ZHANG NL, FU C, LIU TF, et al. A data-driven method for syndrome type identification and classification in traditional Chinese medicine[J]. J Integr Med, 2017, 15(2): 110-123.
- [16] 蔡伟杰, 张晓辉, 朱建秋, 等. 关联规则挖掘综述[J]. 计算机工程, 2001(5): 31-33, 49.
- [17] 杨伟. 有向无圈图(DAGs)的学习算法研究[D]. 北京: 中国人民大学大学, 2012.
- [18] 张连文, 许朝霞, 王忆勤, 等. 隐结构分析与西医疾病的辨证分型(Ⅱ): 综合聚类[J]. 世界科学技术-中医药现代化, 2012, 14(2): 1422-1427.
- [19] 张连文, 袁世宏. 隐结构模型与中医辨证研究(Ⅰ): 隐结构法的基本思想及隐结构分析工具[J]. 北京中医药大学学报, 2006, 29(6): 5.
- [20] YANG W, YI DH, XIE YM, et al. Statistical identification of syndromes feature and structure of disease of western medicine based on general latent structure model [J]. Chin J Integr Med, 2012, 18(11): 850-861.
- [21] 张连文, 傅晨, 刘腾飞, 等. 隐结构分析与西医疾病的辨证分型(Ⅲ): 辨证规则的建立[J]. 世界科学技术-中医药现代化, 2014, 16(4): 723-730.
- [22] 徐孟娇, 金潇, 袁文雯, 等. 基于贝叶斯网络与关联规则探讨下肢动脉硬化闭塞症的中医临床疗效评价条目[J]. 中医杂志, 2023, 64(12): 1218-1224.
- [23] 沈照峰, 刘小娟, 朱磊, 等. 基于复杂网络和隐结构模型的溃疡性结肠炎中医证候要素及辨证规律研究[J]. 中医杂志, 2023, 64(3): 280-287.
- [24] 宋全枚, 张学成, 张永臣, 等. 基于隐结构模型探析后溪穴古代配伍规律[J]. 针刺研究, 2022, 47(5): 459-465.
- [25] 陈丽平, 李建生, 白云苹, 等. 基于隐结构结合灰色关联探讨中医古籍咳嗽病辨治规律[J]. 中国中医基础医学杂志, 2021, 27(5): 777-781, 817.
- [26] 李秦, 李桓, 李松伟. 基于隐结构模型的古代痹证辨证分型及用药规律研究[J]. 中国中药杂志, 2020, 45(19): 4784-4791.
- [27] 李宁, 谢兴文, 陈欣. “肾主骨、生髓”理论在骨质疏松症研究中的应用[J]. 中国中医骨伤科杂志, 2011, 19(4): 70-72.
- [28] 吴国庆, 皮持衡. 基于“肾主骨”理论探讨肾性骨病的中医治疗[J]. 中国中医药现代远程教育, 2020, 18(6): 47-50.

Constructing a Syndrome Differentiation Model by Combining Latent Structure Analysis of Association Rules: An Example of Kidney Yang Deficiency Syndrome in Primary Osteoporosis

LI Beibei^{1,2}, WANG Jing^{3,4}, XIANG Xinghua¹, XU Wenyuan¹, BAI Weiguo¹, LIU Mengyu¹, LI Yukun¹, WU Xiaokun², WANG Yongjun^{3,4}, YANG Wei¹

1. Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, 100700; 2. Department of Mathematics and Science, North China Electric Power University; 3. Institute of Spine Disease, Shanghai University of Traditional Chinese Medicine; 4. Longhua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine

ABSTRACT **Objective** To construct a quantitative differentiation model of traditional Chinese medicine (TCM) syndromes by taking primary osteoporosis (POP) with kidney *yang* deficiency syndrome as an example, and to provide methodological reference for the standardization of syndrome differentiation. **Methods** High-frequency clinical features of POP were screened by descriptive statistical analysis, and strong association features of POP were obtained by association rule algorithm. On this basis, a latent structure (latent tree) model was established through latent structure analysis, and the implicit and explicit variables (features) related to POP with kidney *yang* deficiency syndrome were comprehensively clustered, and the clustering results were interpreted by the indexes of mutual information and cumulative information coverage, to explore the primary and secondary symptoms, and to deduce the categories of POP with kidney *yang* deficiency syndrome based on the probability of the features appearing in the various latent categories. Based on the categories, the clinical feature scores and identification thresholds were calculated, and the syndrome differentiation model of POP with kidney *yang* deficiency was initially constructed by combining the comprehensive judgment rules. Finally, the results of TCM professionals' judgment were used as the gold standard to further evaluate the effectiveness of the model in assisting the syndrome differentiation. **Results** The 32 features strongly associated with POP were obtained, and the Bayes information criterion score of the further constructed latent tree model was -15291.93. Based on the mutual information and the cumulative information coverage, the main symptoms of POP with kidney *yang* deficiency syndrome were bone weakness, fatigue, pale tongue, clear urine, frequent nocturnal urination, cold limbs, thin pulse, white coating, and secondary symptoms were weakness, loss of libido, loose stool, frequent urination, lumbar and knee weakness, and fear of cold. From the probability of the occurrence of each clinical feature in different latent categories of POP with kidney *yang* deficiency syndrome, the state was introduced as S0 category (none/mild kidney *yang* syndrome)/ S1 category (moderate kidney *yang* syndrome)/ S2 category (severe kidney *yang* syndrome). Optimizing the preliminary rules of state identification and refining the state of S1 category, the results showed that among 970 patients with POP, there were 520 patients having no/mild kidney *yang* deficiency syndrome, 224 patients with moderate to mild kidney *yang* deficiency syndrome, 81 patients with moderate to severe kidney *yang* deficiency syndrome, and 145 patients with severe kidney *yang* deficiency syndrome. During the evaluation and validation process, the correct rate of the model assessment index was 0.8835, while the sensitivity was 0.7181, and the specificity was 0.9437. **Conclusion** Combined with the latent structure analysis of the association rule, the syndrome differentiation model for POP with kidney *yang* deficiency could be constructed, and the model shows a good quantitative identification effect, which can provide methodological supports for clinicians to improve the efficiency and accuracy of TCM diagnosis.

Keywords primary osteoporosis; association rules; latent structure (latent tree); kidney *yang* deficiency syndrome; syndrome differentiation model

(收稿日期: 2022-08-21; 修回日期: 2023-10-13)

[编辑: 焦 爽]