上海市 35 岁以上居民慢性病相关行为危险因素关联情况的挖掘*

上海市疾病预防控制中心(200336)　刘晓侠　杨群娣　刘丹妮　苏秋云　郑　杨　施　燕△

【提　要】 **目的**　使用关联规则挖掘上海市≥35 岁常住居民慢性病相关行为危险因素的关联性及关联强度。**方法**　利用上海市 2017 年慢性病及其危险因素监测数据,描述≥35 岁常住居民现在吸烟、危险或有害饮酒、蔬菜和水果摄入不足、红肉摄入过量、身体活动不足和每天睡眠时长不合格 6 项行为危险因素共存情况,采用 R 4.1.2 软件 arules 包 Apriori 算法对 6 项行为危险因素在不同性别和年龄组人群的关联情况进行关联规则分析。**结果**　19578 名居民中至少存在 2 项行为危险因素的占比为 39.4%。支持度超过 5%、置信度超过 50%且提升度最高的关联规则为:男性 35~59 岁组{危险或有害饮酒}→{现在吸烟}(C=82.38%)、男性≥60 岁组{危险或有害饮酒}→{现在吸烟}(C=63.78%)、女性≥60 岁组{蔬菜和水果摄入不足}→{睡眠时长不合格}(C=51.35%)。**结论**　上海市≥35 岁常住居民慢性病行为危险因素共存现象常见且相互间存在关联,不同特征人群的共存关联特点不一,提示在今后的行为危险因素干预项目中,需提高行为危险因素综合干预的意识,根据不同特征人群特点实施有针对性的多项行为危险因素综合干预措施。

【关键词】 慢性病　行为危险因素　关联规则　Apriori 算法

【中图分类号】 R181.3　　【文献标识码】 A　　**DOI**　10.11783/j.issn.1002-3674.2024.01.014

2019 年中国因慢性病导致的死亡人数占到总死亡人数的 88.5%,慢性病防控工作面临巨大的挑战[1]。吸烟、有害饮酒、缺乏体育活动、不健康饮食和长期的睡眠不足是慢性病发生和发展的主要行为危险因素,这些因素在人群中持续流行加大了慢性病的患病风险[2-5]。研究发现,坚持健康行为与降低死亡率和慢性病患病风险息息相关,改变危害健康的行为,能够成功预防 75%以上慢性病的发生,行为危险因素的控制对于预防慢性病控制至关重要[6-8]。

慢性病相关行为危险因素如何共变、聚集、分布以及相互作用在发达国家已经得到了很好的研究[9-11],国内对相关行为危险因素的聚集和相关性关注不多。中国疾病预防控制中心对全国近 5 万名 15~69 岁居民分析了包括吸烟、过量饮酒、蔬菜水果摄入不足、缺乏身体活动及超重或肥胖的流行率、共变异、聚类和独立相关,识别了容易危险因素共存的人群,为开展针对多个危险因素的综合干预提供支持[12]。本研究旨在对上海市慢性病危险因素监测数据进行慢性病相关行为危险因素的共存状况进行统计分析,运用关联规则挖掘慢性病相关主要行为危险因素的关联性及关联强度,为今后上海市居民慢性病相关行为危险因素干预及落实全民健康生活方式提供数据支撑,进而为实现慢性病综合防控相关发展目标而努力[13]。

*基金项目:上海市医学领军人才(2019LJ24);上海市卫生健康委员会卫生行业临床研究专项项目(20204Y0195);上海市“医苑新星”青年医学人才培养-公共卫生领导者项目;上海市公共卫生体系建设三年行动计划学科建设项目(GWV-10.1-XK05)

△通信作者:施燕,E-mail:shiyan@scdc.sh.cn.

资料与方法

1.数据来源

本研究数据来源于 2017 年上海市第四轮慢性病及其危险因素监测项目[14]。该项目通过现场调查方式采集居民慢性病相关行为危险因素(吸烟、饮酒、膳食、运动等)及慢性病(高血压、糖尿病等)患病情况等信息。本研究选取项目中年龄≥35 岁的居民作为研究对象,并选取项目中慢性病相关行为危险因素的条目内容作为数据集。

2.分析指标

根据我国重要慢性病相关危险因素[15]分布情况选择分析指标:现在吸烟、危险或有害饮酒(过去 12 个月)、蔬菜和水果摄入不足、红肉摄入过多、身体活动不足和每天睡眠时长不合格,指标评估标准如下[16]:(1)现在吸烟:调查时吸烟。(2)危险或有害饮酒:过去 12 个月,男性饮酒者平均每天纯酒精摄入量≥41g,女性饮酒者平均每天纯酒精摄入量≥21 g。(3)蔬菜和水果摄入不足:每人每日蔬菜和水果平均摄入量低于 400 g。(4)红肉摄入过多:每人每日猪、牛、羊肉等红肉类食物平均摄入量按生重计超过 100 g。(5)身体活动不足:每周中等强度活动(包括干农活、工作、家务、交通相关的身体活动、休闲性锻炼或运动等)时间不足 150 min 或相当量。(6)每天睡眠时长不合格:成人每天睡眠合格时长为 7~8 h[3],低于 7 h 或高于 8 h 定义为每天睡眠时长不合格。

3.统计分析

对研究对象行为危险因素共存现状进行描述性统计;采用关联规则 Apriori 算法挖掘行为危险因素的关联状况[17]。Apriori 算法是关联规则的基本算法,是数

据挖掘领域十大经典算法之一，它由 Agrawal 和 Srikant 在 1994 年提出[18]，包括寻找频繁项集和依据频繁项集产生关联规则两个步骤，这些规则必须同时满足预先设定的最小支持度和最小置信度的阈值。对于行为危险因素 $\{X\}$ 与 $\{Y\}$ 之间的关联规则 $\{X\}\rightarrow\{Y\}$：

支持度 $S_{\{X\}\rightarrow\{Y\}}=\frac{|T(X\cap Y)|}{|T|}$，即数据集内 $\{X\}$、$\{Y\}$ 同时出现的频率；

置信度 $C_{\{X\}\rightarrow\{Y\}}=\frac{|T(X\cap Y)|}{|T(X)|}$，即数据集内出现 $\{X\}$ 的例数中出现 $\{Y\}$ 的频率；

提升度 $L_{\{X\}\rightarrow\{Y\}}=\frac{C_{\{X\}\rightarrow\{Y\}}}{S_{\{Y\}}}=\frac{S_{\{X\}\rightarrow\{Y\}}}{S_{\{X\}}S_{\{Y\}}}$，即数据集内出现 $\{X\}$ 的例数中出现 $\{Y\}$ 的频率与数据集 $\{Y\}$ 出现的频率的比值。

支持度与置信度分别强调 $\{X\}\rightarrow\{Y\}$ 的发生具有一定强度，提升度则要求此关联下发生的可能性提高了多少倍。

本研究统计分析通过 R 4.1.2 软件实现。通过 R 软件 arules 包 apriori 函数进行关联规则提取，最小支持度设置为 5%、最小置信度设置为 50%，提升度设置为大于 1。

结　果

1.研究对象基本情况

剔除研究指标含缺失值的调查对象后，本研究共纳入研究对象 19578 人。其中：男性 7913 人（40.4%），女性 11665 人（59.6%）；35～59 岁 7131 人（36.4%），≥60 岁 12447 人（63.6%）。

2.行为危险因素基本情况

19578 名研究对象中 6 项行为危险因素占比由高至低依次为：睡眠时长不合格（44.6%）、蔬菜和水果摄入不足（37.0%）、身体活动不足（19.2%）、现在吸烟（16.3%）、红肉摄入过多（10.9%）、危险或有害饮酒（3.7%）。男性与女性各项行为危险因素分布均存在显著性差异（P<0.05），详见表 1。除危险或有害饮酒、蔬菜和水果摄入不足两因素外，35～59 岁与≥60 岁组其余项行为危险因素分布均存在显著性差异（P<0.05），详见表 2。

表 1　不同性别慢性病相关行为危险因素分布情况

行为危险因素	男性（%）	女性（%）	χ^2	P
睡眠时长不合格	43.6	45.4	6.3	0.01
蔬菜和水果摄入不足	41.9	36.1	67.4	<0.01
现在吸烟	39.5	0.7	5180.0	<0.01
身体活动不足	24.7	15.4	260.9	<0.01
红肉摄入过多	14.1	8.7	141.0	<0.01
危险或有害饮酒	8.6	0.3	913.8	<0.01

表 2　不同年龄慢性病相关行为危险因素分布情况

行为危险因素	35～59 岁（%）	≥60 岁（%）	χ^2	P
睡眠时长不合格	38.7	48.1	161.0	<0.01
蔬菜和水果摄入不足	38.1	38.6	0.4	0.53
现在吸烟	18.7	15.1	43.6	<0.01
身体活动不足	17.4	20.1	21.4	<0.01
红肉摄入过多	12.7	9.8	41.3	<0.01
危险或有害饮酒	3.6	3.7	0.1	0.75

研究对象中有 39.4%的人有 2 项及以上行为危险因素。男性中有 55.1%的人有 2 项及以上行为危险因素，女性中有 28.8%的人有 2 项及以上行为危险因素，男性与女性行为危险因素个数存在显著性差异（χ^2=2091.4，P<0.05）。35～59 岁中有 37.6%的人有 2 项及以上行为危险因素，≥60 岁中有 40.5%的人有 2 项及以上行为危险因素，35～59 岁与≥60 岁人群行为危险因素个数存在显著性差异（χ^2=41.0，P<0.05），详见表 3。

表 3　不同性别和年龄的行为危险因素数组合

行为危险因素个数	性别		年龄		合计（%）
	男性（%）	女性（%）	35～59 岁（%）	≥60 岁（%）	
0	13.9	27.4	24.1	20.8	22.0
1	31.0	43.7	38.3	38.8	38.6
2	31.6	23.8	25.4	27.9	27.0
3	17.1	4.7	9.2	10.0	9.7
4	5.3	0.3	2.6	2.2	2.3
5	1.0	0.0	0.4	0.4	0.4
6	0.1	0.0	0.1	0.0	0.0

3.关联规则分析结果

不同性别和年龄组的关联分析结果显示，男性 35～59 岁组有 9 项、男性≥60 岁组有 4 项、女性 35～59 岁组有 0 项和女性≥60 岁组有 1 项满足条件。详见表 4。

男性 35～59 岁组现在吸烟与其余 5 项行为危险因素均存在关联，最流行的行为危险因素共存模式为蔬菜和水果摄入不足和现在吸烟（S=22.57%）；关联规则 $\{$危险或有害饮酒$\}\rightarrow\{$现在吸烟$\}$ 的置信度和提升度最高（S=7.61%，C=82.38%，L=1.69），支持度 7.61%提示有 7.61%的人同时有危险或有害饮酒和现在吸烟行为，置信度 82.38%提示有危险或有害饮酒行为的人中有 82.38%的人同时有现在吸烟的行为，提升度 1.69 表示如果一个人有危险或有害饮酒的行为，那么这个人同时有现在吸烟行为的概率是这个人在没有这个假定条件情况下的 1.69 倍。

男性≥60 岁组现在吸烟与危险或有害饮酒，蔬菜和水果摄入不足与身体活动不足、睡眠时长不合格，睡眠时长不合格与现在吸烟、蔬菜和水果摄入不足之间存在关联，最流行的行为危险因素共存模式为现在吸烟，蔬菜和水果摄入不足和睡眠时长不合格（S=7.87%）；关联规则 $\{$危险或有害饮酒$\}\rightarrow\{$现在吸烟$\}$

的置信度和提升度最高（S = 5.31%，C = 63.78%，L = 1.83）。

女性≥60 岁组蔬菜和水果摄入不足和睡眠时长不合格存在关联且共存比例较高（S = 19.02%）。

表 4　关联规则筛选结果

性别	年龄	前项	后项	支持度（%）	置信度（%）	提升度
男性						
	35～59 岁					
		危险或有害饮酒	现在吸烟	7.61	82.38	1.69
		身体活动不足，睡眠时长不合格	现在吸烟	5.53	56.15	1.15
		蔬菜和水果摄入不足，睡眠时长不合格	现在吸烟	9.09	54.55	1.12
		睡眠时长不合格	现在吸烟	20.11	53.05	1.09
		蔬菜和水果摄入不足，身体活动不足	现在吸烟	6.47	52.94	1.08
		蔬菜和水果摄入不足	现在吸烟	22.57	50.94	1.04
		红肉摄入过多	现在吸烟	8.78	50.43	1.03
		身体活动不足	现在吸烟	12.76	50.37	1.03
		现在吸烟，身体活动不足	蔬菜和水果摄入不足	6.47	50.74	1.15
	≥60 岁					
		危险或有害饮酒	现在吸烟	5.31	63.78	1.83
		身体活动不足，睡眠时长不合格	蔬菜和水果摄入不足	5.96	50.65	1.25
		蔬菜和水果摄入不足，身体活动不足	睡眠时长不合格	5.96	50.97	1.10
		现在吸烟，蔬菜和水果摄入不足	睡眠时长不合格	7.87	50.80	1.09
女性						
	≥60 岁					
		蔬菜和水果摄入不足	睡眠时长不合格	19.02	51.35	1.04

讨　论

关联规则是数据挖掘中关联分析成果的核心体现形式。目前关联规则在慢性病防控领域有一些应用，如分析慢性代谢性疾病对脂肪肝、肥胖对 3 种常见慢病（高血压、冠心病和糖尿病）、慢性病相关危险因素关联模式等[19-23]，分析结果为慢性病筛查和诊断、临床决策等提供了新的思路和一定的理论支持。本研究选取流行率较高的 6 项行为危险因素，对上海市慢性病及其危险因素监测项目中近 2 万名≥35 岁常住居民进行了关联规则分析，其结果通过支持度、置信度和提升度三个统计量进行了直观、清晰的展现。

本研究显示，上海市≥35 岁居民中同时存在 2 种及以上行为危险因素的比例较高（39.4%），与陕西省成年人群（54.58%）[24]、江苏省南京市成年人群（54.97%）[25]和宁夏农村地区 25～74 岁居民（84.82%）[26]一样具有较高的行为危险因素共存水平，与《中国居民营养与慢性病状况报告（2020 年）》[1]中我国居民不健康生活方式广泛存在的结论一致。

本研究显示，不同特征人群行为危险因素的暴露水平、共存水平和关联特点存在差异，危险因素的共存模式呈现多样化和复杂化的特点。提示需对不同特征人群行为危险因素的暴露情况开展有针对性的干预和指导，提高针对多种行为危险因素开展综合干预的意识，从而有效降低行为危险因素的暴露水平及其聚集风险，降低行为危险因素的共同作用对慢性病等疾病发生发展的影响。对于上海市≥35 岁男性人群，具有危险或有害饮酒行为的人更有可能同时存在现在吸烟行为，因此，在控烟干预中联合控酒措施，可以更有效地达到控烟目的。

关联规则分析方法常被应用于对多种慢性病的共病及关联情况的分析，而既往研究常采用枚举的方法对多项危险因素的共存情况进行逐一列举分析。本研究应用关联分析方法帮助研究人员简便、快捷地实现了目标人群行为危险因素共存特点的分析，探索预测了具有特定行为危险因素的人群可能存在的其他危险因素，为针对这一行为危险因素共存模式采取综合干预措施落实全民健康生活方式提供了科学依据。本研究的局限性之一是各组数据集中均获得大量的关联规则，最小支持度和置信度的设置会影响最后结果，需根据数据特点及专业知识进行设置；二是数据集中危险因素指标通过研究对象自报获得，信息可能存在偏倚；三是行为危险因素选择了慢性病危险因素监测项目监测内容中流行率较高的 6 项指标，可能不够全面。

参 考 文 献

[1] 中华人民共和国国务院新闻办公室.《中国居民营养与慢性病状况报告（2020 年）》发布会图文实录（2020-12-23）[2021-11-09] http://www.scio.gov.cn/xwfbh/xwbfbh/wqfbh/42311/44583/wz44585/Document/1695276/1695276.htm.

[2] World Health Organization. Noncommunicable diseases: Overview [2021-11-09] https://www.who.int/health-topics/noncommunicable-diseases#tab=tab_1.

[3] 中华人民共和国国家卫生健康委员会.《健康中国行动（2019-

2030年)》(2019-07-15)[2021-11-9]http://www.nhc.gov.cn/guihuaxxs/s3585u/201907/e9275fb95d5b4295be8308415d4cd1b2.shtml.
[4] 黄瑛,朱娜,秦虹云,等.社区老人睡眠质量与慢性病的相关性分析.国际精神病学杂志,2021,48(4):656-659.
[5] 王丽敏,关云琦. 睡眠状况与主要慢性病患病的关系. 中华流行病学杂志,2020,41(8):1237-1241.
[6] Khaw KT, Wareham N, Bingham S, et al. Combined impact of health behaviours and mortality in men and women: the EPIC-Norfolk prospective population study. PLoS Med,2008, 5(1): e12.
[7] Knoops KTB, de Groot LC, Kromhout D, et al. Mediterranean diet, lifestyle factors, and 10-year mortality in elderly European men and women: the HALE project. JAMA,2004, 292(12): 1433-1439.
[8] 刘晓君.我国成年人健康相关行为现状及其社会生态学因素研究.武汉大学,2020.
[9] Fine LJ, Philogene GS, Gramling R,et al. Prevalence of multiple chronic disease risk factors:2001 National Health Interview Survey. Am J Prev Med,2004,27(2):18-24.
[10] Lowry R, Kann L, Collins JL, et al. The Effect of Socioeconomic Status on Chronic Disease Risk Behaviors Among US Adolescents. JAMA, 1996, 276(10):792-797.
[11] Alamian A, Paradis G. Correlates of Multiple Chronic Disease Behavioral Risk Factors in Canadian Children and Adolescents. Am J Epidemiol, 2009, 170(10):1279-1289.
[12] Yichong L, Mei Z, Yong J, et al. Co-variations and Clustering of Chronic Disease Behavioral Risk Factors in China: China Chronic Disease and Risk Factor Surveillance, 2007. PLoS ONE, 2012, 7(3): e33881.
[13] 吴菲, 郑杨, 程旻娜. 重视危险因素监测,助力慢性病有效防控. 上海预防医学, 2019, 31(2): 81-83.
[14] 陆晔,刘晓侠,朱珍妮,等. 上海市中老年人群优质蛋白质摄入与肌肉衰减症风险的关系研究. 营养学报,2020,42(5):429-434.
[15] 王丽敏,张梅,周脉耕,等.中国慢性病及危险因素监测创新技术体系构建与应用研究.中华流行病学杂志,2021,42(7):1154-1159.
[16] 上海市疾病预防控制中心.上海市慢性病及其危险因素监测报告(2013).上海:上海科学普及出版社,2014:12-14.
[17] 薛薇.R语言数据挖掘.第2版.北京:中国人民大学出版社,2018:299,318-319.
[18] Agrawal R, Srikant R. Fast algorithms for mining association rules [C]//Proc. 20th int. conf. very large data bases, VLDB. 1994, 1215: 487-499.
[19] 管佩霞,李志强,毛倩,等. 基于关联规则的脂肪肝与5种慢性代谢性疾病的关联性分析. 现代预防医学,2021,48(17):3242-3246+3253.
[20] 陈晨,王妮,黄艳群,等. 基于居民健康大数据的肥胖与常见慢病关联规则分析. 北京生物医学工程,2020,39(4):406-411+417.
[21] 朱碧云,王妮,黄艳群,等. 高血压患者合并重大慢性病关联规则分析. 医学信息学杂志,2019,40(11):66-70.
[22] 罗旋,闫晓芳,罗明明,等. 海淀区居民慢性病相关危险因素的关联规则. 中南大学学报(医学版),2017,42(5):570-574.
[23] 郭慧敏,杜军,黄路非. 基于R的Apriori算法在高额住院费用中的应用研究. 中国卫生统计,2017,34(2):315-317.
[24] 王维华,飒日娜,邱琳,等. 陕西省慢性病行为危险因素聚类特征分析. 中国慢性病预防与控制,2021,29(1):18-22+28.
[25] 王巍巍,苏健,周金意,等. 2017-2018年南京市成人居民慢性病相关危险因素. 中华疾病控制杂志,2021,25(11):1264-1268,1275.
[26] 常晓玉,张田敬,李博宇,等. 宁夏农村居民慢性病相关行为生活方式分布状况研究. 宁夏医学杂志,2016,38(7):632-635.

(责任编辑:郭海强)

(上接第67页)

[13] Osone A, Arai R, Hakamada R, et al. Cognitive and brain reserve in conversion and reversion in patients with mild cognitive impairment over 12 months of follow-up. Journal of Clinical and Experimental Neuropsychology. 2016, 38(10): 1084-1093.
[14] Proust C, Jacqmin-Gadda H, Taylor JM, et al. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. Biometrics, 2006, 62(4): 1014-1024.
[15] Weiner M, Aisen P, Jack C, et al. The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans. Alzheimers & Dementia, 2010, 6(3): 202-211.e7.

(责任编辑:邓 妍)