中青年心肌梗死患者院前症状及就医情况的文本挖掘

黄晶晶^{1,2}, 黄素芳¹, 王荃^{1,2}, 刘雨晨^{1,2}, 张可心^{1,2}, 盛晓萱^{1,2}, 刘诗雅^{1,2}

摘要:目的 了解中青年急性心肌梗死患者的院前症状和就医特征,为医护人员开展针对性健康教育以改善患者就医延迟提供参考。方法 根据纳入及排除标准,从医院的医疗大数据平台提取 200 例中青年急性心肌梗死患者的电子病历数据,对其电子病历中的非结构化文本信息进行文本挖掘,使用词频分析和可视化方法分析患者院前症状和就医情况。结果 胸痛和胸闷是中青年急性心肌梗死患者的典型和特异症状,大汗、乏力、心慌等是次常见和次强相关的院前症状。148 例患者选择当地就诊后寻求进一步诊治而转诊上级医院。结论 中青年急性心肌梗死患者就医前有不同的典型症状或非典型症状,患者就医情况与病情有关。医疗大数据平台在操作便利性、数据获取能力方面展现出独特优势,但数据质量需提高。

关键词:中青年; 急性心肌梗死; 院前症状; 就医延迟; 医疗大数据平台; 电子病历; 文本数据; 文本挖掘

中图分类号: R473.5 **DOI:** 10.3870/j.issn.1001-4152.2024.18.028

Text mining of pre-hospital symptoms and medical-seeking behaviors in young and middle-aged patients with myocardial infarction

Huang Jingjing, Huang Sufang, Wang Quan, Liu Yuchen, Zhang Kexin, Sheng Xiaoxuan, Liu Shiya, Department of Emergency, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

Abstract: **Objective** To understand the pre-hospital symptoms and medical-seeking characteristics of young and middle-aged patients with acute myocardial infarction, so as to provide a reference for medical staff conducting targeted health education to improve patient delays in seeking medical attention. **Methods** Adhering to inclusion and exclusion criteria, electronic medical records of 200 young and middle-aged acute myocardial infarction patients were extracted from the hospital's medical big data platform. Text mining was performed on the unstructured textual information within the electronic medical records, then frequency analysis and visualization techniques were employed to analyze patients' pre-hospital symptoms and medical-seeking behavior. **Results** Chest pain and tightness were identified as the typical and specific symptoms of young and middle-aged patients with acute myocardial infarction, while profuse sweating, fatigue, and palpitations were among the less common but strongly related pre-hospital symptoms. A total of 148 patients sought local medical attention and were subsequently referred to higher-level hospitals for further diagnosis and treatment. **Conclusion** The young and middle-aged patients with acute myocardial infarction exhibit varying degrees of typical and atypical symptoms prior to seeking medical care, and their medical-seeking behavior is related to the severity of their condition. The medical big data platform demonstrates unique advantages in terms of operational convenience and data acquisition capabilities, however, the quality of data still requires improvement.

Keywords: young and middle-aged; acute myocardial infarction; pre-hospital symptoms; delay in medical-seeking; medical big data platform; electronic medical records; textual data; text mining

医疗行业的快速发展,产生了海量的医疗数据,亟需专业平台进行有效管理和利用^[1]。医疗大数据平台的建设为临床研究和患者管理提供了前所未有的机遇。医疗大数据平台数据来源于以患者为中心的不同信息系统,包括住院信息系统、门诊及住院电子病历、实验室信息管理系统、护理信息系统等^[2]。文本数据作为医疗大数据的重要组成部分,为患者病程的变化提供了客观的描述。这些数据不仅映射了患者的健康状态,也辅助医疗专业人员监测治疗成效、预测疾病走向,对于理解疾病模式以及提升护理质量具有不可替代的价值^[3]。急性心肌梗死(Acute Myocardial Infarction, AMI)是一种危及生命的心血管疾病,随着生活节奏变快和不良生活方式的增加,AMI发病率逐渐呈年轻化趋势^[4]。及时有效地诊断和治疗 AMI,是减少

AMI 危害和负担的关键。有研究显示,及时再灌注治疗延迟与患者病死率增加相关^[5]。据文献报道,我国 AMI 患者症状发作后到院时间和再灌注治疗时间均存在普遍延迟^[6]。目前对于 AMI 患者院前症状和就医行为研究,多是基于问卷调查或医疗记录的数据,由人工收集而来,收集的数据可能存在一定偏差和误差,不能完全反映患者真实症状和就医情况。本研究基于医疗大数据平台,对中青年 AMI 患者电子病历非结构化文本信息进行文本挖掘,回顾性分析患者院前症状和就医情况,探讨不同症状之间的关联程度和变化趋势,揭示 AMI 患者院前症状和就医行为特征,为医护人员开展针对性健康教育以改善患者就医延迟提供参考。

1 资料与方法

1.1 资料来源 2023 年 12 月,采用随机抽样方法提取我院心内科住院的中青年 AMI 患者的电子病历。入选标准为:①入院诊断为 AMI、AMI 待排、心绞痛;②年龄 18~59 岁;③首次入院。排除标准:并存发热、低蛋白血症、重度营养不良、肝肾功能不全、血液疾病、恶性肿瘤、终末期疾病。从我院科研大数据平台提取

作者单位:1. 华中科技大学同济医学院附属同济医院急诊科

(湖北 武汉, 430030); 2. 华中科技大学同济医学院护理学院

黄晶晶:女,硕士在读,学生, huangjingjing@hust.edu.cn

通信作者:黄素芳, huangsufang@tjh.tjmu.edu.cn

收稿:2024-03-29;修回:2024-06-01

358 例中青年 AMI 患者,通过人工方法逐一复查,排除因系统疏漏而纳入的非首次入院患者,排除他院经冠状动脉介入术后患者。最终纳入 200 例中青年 AMI 患者的电子病历资料进行文本挖掘。

1.2 方法

1.2.1 文本数据的提取 提取 2016 年 1 月 1 日至 2023 年 12 月 1 日住院中青年 AMI 患者电子病历的文本数据。由研究团队的护理硕士研究生负责进行文本数据提取,医院信息科及科研大数据平台工作人员作为技术顾问。从医院科研大数据平台提取数据,研究团队在平台上创建专门的科研项目,并根据纳入及排除标准筛选符合条件的患者。提取的数据资料包括:患者基本信息(性别、年龄、文化程度、职业、医疗付款方式)、现病史、个人史(抽烟史、饮酒史、毒品接触史)和既往疾病史等。数据导出申请经过信息科的审核和批准后,数据被安全地存储在受到严格监管的云桌面。

1.2.2 文本数据的清洗和预处理 关注患者现病史中的主诉部分,提取首次发病时间、诱因、症状、部位、伴随症状、持续时间、处理情况和就医情况等关键信息,去除患者入院辅助检查和模板套话等无关文本数据。通过此过程,提取关于患者发病症状和就医情况的高频关键词(频次 ≥ 10)。整理后,统一书写表达格式,以排除文字歧义和书写风格差异造成的影响。此外,逐一核查患者基本信息、既往史和个人史,以避免出现文字前后矛盾的情况。

1.2.3 文本数据的分析方法 ①分词和词频分析。分词是指将患者的现病史记录切分成单个词语,以便于对文本进行分析和处理。本研究通过 ROSTCM 6 软件^[7]调整分词自定义词表,以有效识别和切分出患者的现病史记录中的医学术语和症状描述。词频分析法是一种对文本数据进行统计和分析的方法,它可以计算文本中每个词或短语出现的次数,以反映文本的主题和特征^[8]。本研究使用 ROSTCM6 软件对患者的症状进行词频分析,统计出不同症状的出现频数,使用 R-Studio 软件制作患者院前症状和就医情况词云图。②共现网络分析。共现网络分析是一种利用网络图来展示和分析数据中不同元素之间的共现关系的方法^[9]。其中节点代表分析单元,边代表共现关系,节点和边的大小、颜色、形状等属性可以用来表示不同的特征,如频次、强度、聚类等。使用 Gephi 软件绘制出共现网络图,分析不同症状之间的关联程度和变化趋势。

2 结果

2.1 中青年 AMI 患者一般资料 纳入患者 200 例,男 158 例,女 42 例;年龄:18~<40 岁 28 例,40~<50 岁 123 例,50~59 岁 49 例。文化程度:文盲 11 例,小学 43 例,初中 63 例,高中 43 例,大专及以上 40 例。不良生活方式:吸烟 59 例,饮酒 56 例。手术史 37 例。职业:公务员 11 例,专业技术人员 6 例,职员 29 例,企业管理人员 2

例,工人 15 例,农民 25 例,自由职业者 3 例,个体经营者 3 例,无业人员 15 例,离退休人员 2 例,其他 89 例。并存疾病:高血压 80 例,糖尿病 24 例,高脂血症 12 例,脑梗死 7 例,其他 5 例。冠心病史 5 例。医疗付款方式:城镇职工基本医疗保险 47 例,城镇居民基本医疗保险 31 例,新型农村合作医疗 33 例,自费 3 例,贫困救助 3 例,公费 5 例,其他 78 例。

2.2 中青年 AMI 患者院前症状和就医情况的高频关键词 见表 1。患者首次发病时多选择未就医($n=93$),或选择当地医院就医($n=77$)和直接来我院就医($n=30$)。未就医患者当症状加重或反复发作后,选择当地就医($n=71$)或来我院就医($n=22$)。由于当地医院治疗有限,为求进一步诊治,患者前往我院就医($n=148$)。

表 1 中青年 AMI 患者院前症状和就医情况高频关键词

项目	关键词	频次
首次发病时间	1 年前	15
	6 个月前	13
	1 个月前	13
	2 周前	10
	症状及伴随症状	胸痛
	胸闷	140
	疼痛	74
	大汗	61
	胸前不适	51
	伴心慌	39
	症状加重	36
	再次发作	29
	反复发作	27
	乏力	25
	压榨性	25
	心悸	22
	闷痛	19
	呼吸困难	18
	呕吐	18
	头晕	17
	恶心	14
	气促	10
诱发原因	无明显诱因	70
	活动后	36
	劳累	20
	饮酒后	18
	夜间睡眠	18
不适部位	快走	11
	心前区	61
	胸骨后	29
	胸前区	19
	左侧胸部	18
	胸骨	15
	放射性	15
	背部	14
持续时间	剑突下	13
	数分钟	70
	十余分钟	35
院前处理情况	30 min	16
	未处理	38
	速效救心丸	13
就医情况	当地就医	77
	未就医	93
	我院就医	30
	转诊就医	148

标准化的病历模板可能会引入冗余信息,对数据分析造成干扰。为了解决这些问题,在数据收集、处理和分析的每个阶段实施严格的质量控制措施。这包括但不限于对病历信息的逐项核查,对文本数据的标准化处理,以及定期审查和更新数据管理流程。⑤分析方法复杂。由于对大数据隐私性和保密性要求,医疗数据的导出受限于指定的虚拟云桌面,该云桌面对联网和数据转移有所限制。这些限制增加了在离线环境下进行常规数据分析的复杂度。

3.3 需加强数据质量把控 为确保文本数据的高质量,对数据源进行质量控制至关重要。为应对医疗数据收集中的挑战,制订标准化流程刻不容缓^[22]。这包括数据的录入、处理、审核的标准操作流程,确保每个步骤都遵循明确的指导原则和质量标准。培训和教育方面,医务人员作为数据录入的第一责任人,统一职前或者在职培训和教育是保证数据质量的关键。此外,建立持续的监督和评估机制,对数据管理流程进行定期审查和更新^[23],以确保数据质量持续改进。

4 结论

本研究基于医疗大数据平台,回顾性分析中青年 AMI 患者非结构化的文本数据,探讨院前不同症状之间的关联及就医情况,发现 AMI 患者在就医前不同程度的典型及非典型症状(如乏力、大汗、恶心呕吐等),非典型症状可能是中青年 AMI 的早期征兆或警示信号。多数中青年 AMI 患者选择当地就诊,而后为寻求进一步诊治,转诊到上级医院,有可能延误治疗。医疗大数据平台在操作便利性方面展现出独特优势,但数据质量仍需关注。数据一致性、歧义性、文本质量及标准化问题,以及病历模板使用问题,需要通过标准化流程和质量控制来解决。

参考文献

- [1] 王觅也,郑涛,李楠,等. 医疗大数据集成及应用平台体系构建[J]. 医学信息学杂志,2019,40(8):37-42.
- [2] 徐玲玲,王昊,林琳,等. 某三甲医院科研大数据平台的临床研究管理要点探析[J]. 中国医学伦理学,2023,36(7):749-753.
- [3] Ford E, Carroll J A, Smith H E, et al. Extracting information from the text of electronic medical records to improve case detection:a systematic review[J]. J Am Med Inform Assoc,2016,23(5):1007-1015.
- [4] Andersson C, Vasan R S. Epidemiology of cardiovascular disease in young individuals[J]. Nat Rev Cardiol,2018,15(4):230-240.
- [5] De Luca G, Suryapranata H, Ottervanger J P, et al. Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: every minute of delay counts[J]. Circulation,2004,109(10):1223-1225.
- [6] 国家心血管病医疗质量控制中心专家委员会冠心病专家工作组. 2019 年中国成人急性 ST 段抬高型心肌梗死医疗质量控制报告[J]. 中国循环杂志,2020,35(4):313-325.
- [7] 阎思宇,李绪辉,陈沐坤,等. 面向真实世界的知识挖掘与知识图谱补全研究(二):非结构化电子病历信息抽取方法及进展[J]. 医学新知,2023,33(5):358-365.
- [8] 丁鑫. 云计算在公共图书馆服务中的应用热点分析:基于词频的统计分析[J]. 四川图书馆学报,2015(1):50-54.
- [9] 钟伟金,李佳. 共词分析法研究(一):共词分析的过程与方式[J]. 情报杂志,2008(5):70-72.
- [10] 高瑜,闫旭,张会君. 急性心肌梗死患者症状典型性与院前延误的关系[J]. 中国老年学杂志,2014,34(23):6592-6593.
- [11] Gokhroo R K, Ranwa B L, Kishor K, et al. Sweating:a specific predictor of ST-segment elevation myocardial infarction among the symptoms of acute coronary syndrome: Sweating In Myocardial Infarction (SWIMI) Study Group[J]. Clin Cardiol,2016,39(2):90-95.
- [12] Blakeman J R, Stapleton S J. An integrative review of fatigue experienced by women before and during myocardial infarction[J]. J Clin Nurs,2018,27(5-6):906-916.
- [13] 刘锦秀,王晓伟,张同乐,等. 青年社会体检者、医务工作者及心肌梗死患者心血管疾病高危因素的对比研究[J]. 中华临床医师杂志(电子版),2023,17(3):249-254.
- [14] 黎晓艳,华先妹,陈爱英,等. 社区主要不良心脏事件高危者院前延迟行为意向调查[J]. 护理学杂志,2022,37(7):82-85.
- [15] 国家卫生健康委. 2021 年我国卫生健康事业发展统计公报[EB/OL]. (2022-07-12)[2023-12-23]. https://www.gov.cn/xinwen/2022-7/12/content_5700670.htm.
- [16] 罗素新,袁霄,夏勇. 从胸痛中心建设看非 PCI 医院在构建 ACS 区域协同救治体系中的作用[J]. 西部医学,2017,29(3):297-300.
- [17] 王明明,邹圣强,胡严严. 急性冠脉综合征患者家庭-社区-医院协同化救治的效果[J]. 护理学杂志,2021,36(7):22-25.
- [18] 罗浩,汪鹏,赵浩宇. 基于医疗大数据的可视化分析与应用[J]. 中国医疗设备,2020,35(11):122-124,128.
- [19] 卫荣,陆亮,钱步月,等. 基于大数据的院级临床科研平台建设与实践[J]. 中国数字医学,2019,14(11):91-93,96.
- [20] Lyons A M, Dimas J, Richardson S J, et al. Assessing EHR data for use in clinical improvement and research[J]. Am J Nurs,2022,122(6):32.
- [21] Kim M K, Roupheal C, McMichael J, et al. Challenges in and opportunities for electronic health record-based data analysis and interpretation[J]. Gut Liver,2024,18(2):201-208.
- [22] D'Souza M, Papadopoulou A, Girardey C, et al. Standardization and digitization of clinical data in multiple sclerosis[J]. Nat Rev Neurol,2021,17(2):119-125.
- [23] 夏丽莉,王荣,陈鹏,等. 基于信息平台的护理质量指标实时监测数据库的建设[J]. 护理学杂志,2019,34(19):5-7.