在线健康社区中睡眠障碍疾病描述文本主题特征研究*

庞盼杏[1] 何彩荣[1] 张 磊[2] 陈景信[1] 石荣丽[1,3] 徐中岳[1,3] 翁开源[1]

([1] 广东药科大学医药商学院 广州 510006 [2] 华南理工大学医院 广州 510655
[3] 国家药品监督管理局药物警戒技术研究与评价重点实验室 广州 510006)

〔摘要〕 **目的/意义** 通过挖掘睡眠障碍疾病描述文本，深入了解睡眠障碍线上问诊的现状与睡眠障碍用户的在线问诊信息主题特征。**方法/过程** 以“好大夫在线”平台为数据源，利用网络爬虫获取睡眠障碍相关医患信息，使用隐含狄利克雷分布模型识别患者疾病描述的主题。**结果/结论** 睡眠障碍涉及科室较分散、治疗方式以药物为主，线上问诊能改善83.2%患者病情。用户疾病描述主题包括用药情况与咨询、外界环境、症状描述、代问与病因。建议平台与医生关注患者用药预后情况、心理健康状况，注重共病科普工作。

〔关键词〕 在线问诊；睡眠障碍；隐含狄利克雷分布模型；主题特征
〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2024.10.011

Study on the Thematic Characteristics of Sleep Disorders Disease Description Texts in Online Health Communities

PANG Panxing[1], HE Cairong[1], ZHANG Lei[2], CHEN Jingxin[1], SHI Rongli[1,3], XU Zhongyue[1,3], WENG Kaiyuan[1]
[1] School of Pharmaceutical Business, Guangdong Pharmaceutical University, Guangzhou 510006, China; [2] South China University of Technology, Guangzhou 510655, China; [3] NMPA Key Laboratory for Technology Research and Evaluation of Pharmacovigilance, Guangzhou 510006, China

〔**Abstract**〕 **Purpose/Significance** Disease description texts are analyzed to reach a deeper understanding of the current status of online consultations for sleep disorders and the thematic characteristics of users with sleep disorders. **Method/Process** Data about sleep disorders from “haodf.com” website is collected by using a web crawler. Furthermore, the main themes about patients’ description are identified by the latent Dirichlet allocation (LDA) model. **Result/Conclusion** The departments of sleep disorders are more dispersed, and the main treatment is drug therapy. Online consultations could improve 83.2% of patients’ condition. The themes of patients’s disease descriptions include medication and consultation, external environment, description of symptoms, surrogate questions and causes. It is suggested that the platform and doctors should pay attention to the prognosis of patients’ medication and mental health status, and pay attention to the popularization of comorbidities.

〔**Keywords**〕 online consultation; sleep disorders; latent Dirichlet allocation (LDA) model; characteristics of themes

〔修回日期〕 2024-03-07

〔作者简介〕 庞盼杏，硕士研究生，发表论文1篇；通信作者：徐中岳，博士，副教授，硕士生导师。

〔基金项目〕 广东省哲学社会科学规划项目（项目编号：GD21SQGL01）；广东省广州市哲学社科“十四五”规划课题（项目编号：2023GZGJ67）。

1 引言

近年来越来越多的患者通过在线医疗平台看病就医。睡眠障碍是常见疾病，影响身心健康。睡眠障碍是睡眠的时间和质量异常、睡眠中出现异常行为或睡眠和觉醒节律性交替出现紊乱的表现[1]。目前，我国有 3 亿多人面临睡眠障碍及相关问题[2]。研究[3-4]表明睡眠障碍与多种疾病相关，例如睡眠障碍可引起糖代谢异常以及加剧认知功能受损。调查[5]显示在线医疗健康服务用户非常关注失眠和精神压力，其关注度排第 4 位和第 5 位。在线医疗平台是聚集患者疾病数据的健康信息服务平台，部分问诊记录公开，不仅给予其他用户就诊经验参考，同时也为睡眠障碍相关研究提供大量数据。然而睡眠障碍患者的疾病描述属于大段非结构化文本，从中难以快速提取用户个体疾病特征。本研究以在线健康社区的医患信息作为数据源，分析线上睡眠障碍健康服务现状。基于在线问诊睡眠障碍患者的疾病描述，运用隐含狄利克雷分布（latent Dirichlet allocation，LDA）模型分析用户特征，挖掘在线问诊用户的潜在需求，为在线医疗平台完善相关服务提供参考。

2 基于 LDA 模型的在线健康社区主题特征分析

LDA 是主题建模的流行方法之一，每个主题中概率最高的单词通常可以表示主题内容[6]，近年来多用于从无结构的网络文本中挖掘用户信息需求。从疾病种类角度，陆泉等[7]发现肿瘤患者健康信息需求主要集中在治疗、病理及病因、检查、术后、预防等方面；于本海等[8]爬取“百度痛风吧”帖子内容研究痛风患者信息交流的主题特征，发现交流内容集中在病理知识、疾病诊疗、药物治疗、情感支持、生活习惯等方面；余佳琪等[9]爬取“甜蜜家园”用户评论内容，对糖尿病患者进行主题识别、探究其情感变化，指出除疾病知识等常见话题外，患者还特别关注血糖控制及血糖仪产品。从特殊人群角度，刘冰等[10]关注备孕期女性健康信息需求，爬取“妈妈网”论坛相关内容进行主题分析与情感分析，发现女性身份转变后信息需求多样化，情感、情绪也呈现复杂性、动态性。相较而言，国外疾病类型研究更加多样，样本也更丰富，Gabriela G 等[11]采集全球 98 个网站的 762 名独特用户的帖子研究女性压力性尿失禁交流主题。Li Y 等[12]利用 LDA 处理接受膀胱癌手术患者自由文本并讨论患者手术前后个人目标的变化，发现患者目标从手术和康复转变为恢复身体和工作、享受生活和更加珍惜家人和朋友。Soowon P 等[13]利用主题分析挖掘在线健康社区有关精神障碍的咨询信息，发现用户还比较关注精神障碍的实用信息如福利待遇、社会适应等。近年来国内利用 LDA 主题分析开展的医疗健康相关研究有疾病细化的趋势，主要用于挖掘用户健康信息需求。国外研究则更关注阶段性主题变化，研究目的包括挖掘个人目标、用户关注点、加强对某种疾病的控制等。国内外研究以患-患为代表的网络问答社区为主，缺少对以医-患为代表的健康社区的研究。

3 资料来源与方法

以“好大夫在线”平台作为数据源，整理睡眠障碍医患信息，对患者疾病描述进行主题分析。主题分析是指对文本、语言材料或数据进行分析和理解，通过无监督方式提取语义，保证主题提取的相对客观性与效率[6]。常见主题分析方法包括词频统计分析、主题模型、情感分析等。本研究使用 LDA 主题模型，可帮助确定疾病描述中的重要词语和短语，并将其归类到不同主题中，从而读取患者的信息主题特征，相较于其他主题模型具有语句粒度更精细、研究层次更丰富的特点。

3.1 数据获取

采用 Python 爬取“好大夫在线”平台 2018—2023 年睡眠障碍数据，包括睡眠障碍医

生信息、睡眠障碍患者诊后评价、睡眠障碍患者问诊记录3 部分。爬取医生信息 1 211 条，包括医生姓名、职称、医院、科室、病友推荐度。选取 20 名医生，条件限制为来自三甲医院、具有主治医师及以上职称。针对病友推荐度为 3. 5 以上的医生，爬取其患者诊后评价 2 000 条，包括疗效满意度、治疗方式、态度满意度、目前病情状态。爬取问诊记录 9 329 条，问诊记录数量对应患者数量，爬取信息包括用户性别、年龄、疾病描述等。

3. 2 数据处理

对睡眠障碍医生信息、患者诊后评价、问诊记录进行统计分析。对问诊记录中的疾病描述手动剔除与睡眠障碍无关、重复、异常、信息不全的数据，最终得到有效数据 4 099 条。利用 Python 对有效数据进一步清洗。首先，去除广告标识、无效链接和地址等无用字符，将英文大写字母转为小写、中文繁体字转为简体字；其次，通过文本去重和机械压缩删除冗余，再剔除缺乏实际作用的短文本；然后，利用 jieba 分词库，结合搜狗语料库的“医学词语大全”以及自定义睡眠障碍语料词库如“睡眠障碍”“入睡困难”“不寐症”“美时玉”“思诺思”“佐匹克隆”等形成用户分词词典；最后，对哈尔滨工业大学停用词典、中文停用词典、百度停用词典进行合并整理，根据文本分词效果添加词语形成自定义停用词词典，如“此外”“所以”“当然”“您好”等无意义高频词。经数据处理形成后续 LDA 建模的基础语料库。

3. 3 主题提取与分析

调用 Python 3. 2 中的 sklearn 库建立 LDA 模型，通过计算主题一致性确定最终主题数量，以最终主题数量使用 Python 中的 pyLDAvis 库进行可视化模型结果展示，本研究框架，见图 1。

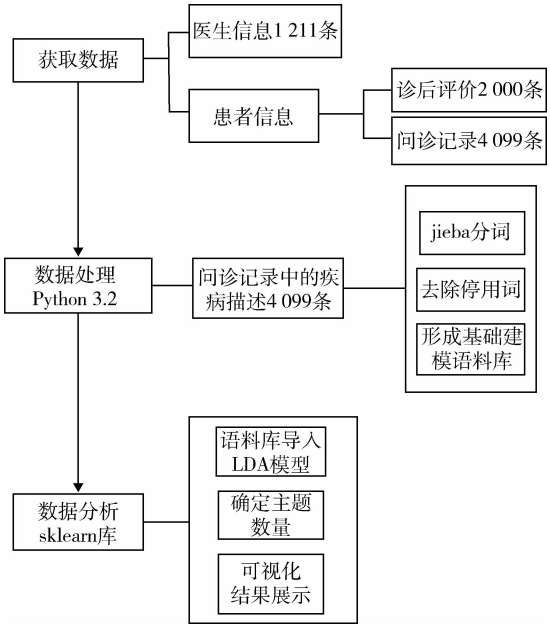

图 1 数据挖掘研究框架

4 结果

4. 1 医生基本情况

1 211 名医生中 281 名未开通线上问诊，930 名可提供在线咨询。科室以神经科（29. 14%）、精神科（21. 29%）、心理咨询科（11. 51%）、中医科（4. 73%）为主。将睡眠医学中心、睡眠障碍科等科室合计，属于睡眠医学的科室仅占 2. 58%，设立睡眠专科的医院集中于一线、二线城市。患者对医生的推荐度最高为 5. 0，最低为 2. 4，其中推荐度高于 3. 7 的医生占比仅为 2. 8%。

4. 2 患者诊后评价情况

2 000 条患者诊后评价中 70. 70% 的患者表示对疗效很满意，76. 90% 的患者表示对医生服务态度很满意。大多数医生采取药物治疗，同一名医生的治疗方案差别不大，例如所调查的神经科医生采取

的治疗方式 70% 以上为药物治疗。83.20% 的患者表示线上问诊后病情有好转，1.30% 的患者表示病情未得到改善甚至加重。

4.3 LDA 最佳主题个数确定

将经过数据清洗后的患者疾病描述文本导入 LDA 模型进行主题一致性测试。主题困惑度与主题一致性是衡量主题质量的常用方法，有研究[14]认为在确定主题数量时一致性指标更科学，且在近两年的相关研究中热度有所提升[6]。一致性表示主题下词语关联的紧密程度，得分越高说明模型拟合得越好。主题数量设为 4 时主题模型一致性得分最高，见图 2。表示主题数量为 4 时包含信息足够多、主题效果较好。结合主题可视化图形变化与人工评价，确定最终主题数量为 4。

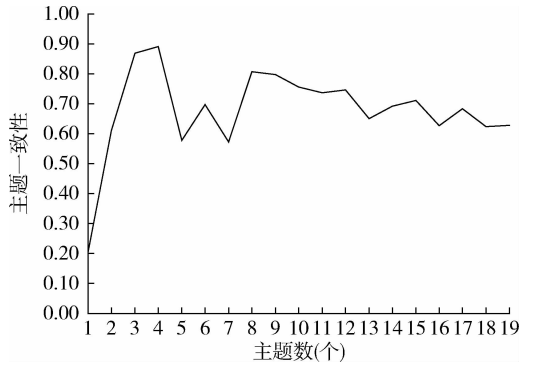

图 2 主题一致性评价

4.4 主题可视化

LDA 主题模型的气泡代表不同主题，气泡间距离代表主题相关度，气泡大小代表主题占数据集的比例。气泡之间无交叉时聚类效果最好。LDA 模型将患者疾病描述文本划分为 4 个主题，见图 3。

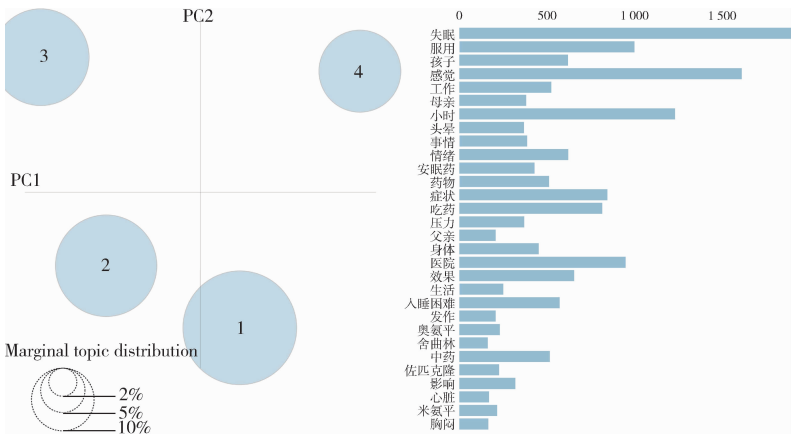

图 3 LDA 模型可视化结果

每个主题选取权重靠前的 25 个高频词，根据每个主题词对应的内容命名主题：用户疾病描述文本中用药情况与咨询主题占比最高，达 32.9%；其次是外界环境对睡眠障碍患者的影响，占比 26.2%；症状描述占比 23.8%，比代问与病因略高，见表 1。

表 1 LDA 主题情况分析

序号	主题名称	主题词	主题占比情况(%)	示例
1	用药情况与咨询	失眠、服用、小时、效果、药物、安眠药、吃药、中药、入睡困难、医院、用药、情况、佐匹克隆、米氮平、盐酸、时间、右佐匹克隆、氯硝西泮、黛力新、胶囊、副作用、阿普唑仑、精神、艾司唑仑、帕罗西汀	32.9	周一复查的，麻烦刘医生看一下报告，是否可以继续吃药

续表 1

序号	主题名称	主题词	主题占比情况（%）	示例
2	外界环境	感觉、失眠、工作、情绪、事情、小时、压力、情况、状态、影响、时间、生活、质量、入睡困难、心情、精神、脑子、身体、症状、心理、睡不好、原因、床上、记忆力、兴趣	26. 2	因工作引起的焦虑抑郁，入睡困难，失眠，心悸，情绪非常低落
3	症状描述	感觉、症状、医院、头晕、身体、发作、头痛、心脏、胸闷、中药、无力、睡眠障碍、血压、精神、全身、焦虑症、躯体、眼睛、吃药、恶心、住院、浑身、情况、气短、脑子	23. 8	我似乎有些躯体反应。比如说，会吐。睡眠状态依旧不好，深睡眠的时间比较短，如果做梦了以后，醒来脑鸣的现象还是有。平时耳鸣是嘶嘶嘶的声音，有时候还有别的声音。很难感知到幸福与愉悦。我比较担心自己，想问问您，这些问题是否能解决
4	代问与病因	孩子、母亲、吃药、父亲、医院、奥氮平、舍曲林、情况、情绪、学校、病情、西酞普兰、睡眠不好、上学、药量、调整、回家、月份、老师、咨询、住院、状态、心理、成绩、电话	17. 1	我母亲 8 个月前帮我带孩子，每天很辛苦，半年前查出肝功能指标异常，经常怀疑自己得了绝症，说是胃部难受，失眠一个月左右，一个月前开始神经异常，自残自杀，在建德第四医院诊断为癔症，住院半个月，吃药维持，出院后继续吃药，仍然睡眠不好，情绪低落，有时难受得控制不住

5 讨论

5. 1 睡眠障碍患者用药咨询特征

主题 1 在所有主题中占比最高，主题词涉及多种药物，其中佐匹克隆、右佐匹克隆属于镇静催眠药物；米氮平、帕罗西汀属于抗抑郁药物；氯硝西泮、黛力新属于抗焦虑药物。部分患者反映服用药物后伴随不良反应，如“患有乙肝，服药后胃疼呕吐难受，还是睡不着”“最近白天特别困，起不来，想问药物是否需要调整”等。“服用”“吃药”“用药”等词反映出患者存在用药咨询现象。治疗方式以药物治疗为主，药物对患者生活质量影响很大，一部分患者因为药物副作用希望调整药物从而减少对生活的影响，另一部分患者则希望不再依赖用药。涉及中药时不少患者使用“浮针”治疗，许多评论反映采取“浮针”治疗结合睡眠认知行为疗法能够停止使用安眠药。这说明现阶段部分睡眠障碍患者已经接受中医治疗睡眠障碍。

5. 2 睡眠障碍患者心理咨询特征与代问特征

社会压力与睡眠障碍有着双向关系，部分患者表示因为外界环境变动导致焦虑、烦躁等情绪进而睡眠质量差，也有患者表示因为失眠感到生活质量下降、工作效率变低。在主题 2 中出现“工作”“压力”“生活”“心情”“心理”等主题词说明用户可能因为外界环境变化直接或间接导致睡眠问题。加之用户除描述疾病本身外更多会描述自身病因，如在经历某些事件之后开始出现睡眠问题。因此需要医生特别关注表现出较大压力的群体，在一定心理治疗的基础上引导患者表达，利用叙事疗法缓解患者病痛[15]。另外，主题 4 中出现“孩子”“母亲”“父亲”等词，用户会代替直系亲属，特别是儿童、老人等弱势群体进行互联网问诊，如“我母亲自更年期开始失眠……”“孩子最近比以前状况要好一些但是还会偶尔出现……”等。对第三人称问诊进行统计，代问占比大于 7%，进一步缓解了弱势群体在线问诊的使用障碍[16]。

5. 3 睡眠障碍患者合并其他疾病特征

通过主题 3 的挖掘，发现部分睡眠障碍患者表示患有高血压、肠胃疾病、尿频等疾病，部分患者患有抑郁症、焦虑症或躁狂症等心理疾病。研究发现高血压、糖尿病、焦虑或抑郁是引发我国老年人失眠的主要危险因素[17]。另外，功能性胃肠病患者常伴有睡眠障碍，而睡眠障碍已被证实是焦虑抑郁躯体化的表现[18]。对在线问诊抑郁症患者划分群组时发现部分群组有严重睡眠问题[19]。参与线上问诊

是积极寻求治疗的第 1 步，提高对睡眠障碍与合并疾病的认识也很重要。值得注意的是，很多其他有关疾病交流研究主题提到预防，但睡眠障碍患者疾病描述中却很少提及，这说明用户对睡眠障碍预防认识不足，睡眠障碍的防控工作任重而道远。

6 结语

通过对医患信息的挖掘，从医生信息来看，目前睡眠障碍诊疗相关科室分散，以适应睡眠疾病的交叉性。从患者信息来看，线上问诊能够帮助大部分睡眠障碍患者改善睡眠问题，很大程度缓解了线下问诊挂号难、距离远、成本高等问题。进一步对睡眠障碍疾病描述进行主题分析，发现睡眠障碍患者存在问诊用药咨询、心理咨询与代问、合并其他疾病等特征。为优化睡眠障碍患者在线问诊服务，本研究提出以下建议。一是针对药物治疗影响患者生活质量的情况，平台与医生应推广非药物治疗，尤其是睡眠认知行为疗法，同时平台应鼓励医生随访患者预后状况，及时调整用药方案。二是针对患者代问情况，平台应优化 App 问诊流程，设计简洁问诊界面，提供温馨提示与引导，简化问诊流程，便于弱势群体自行线上问诊。三是针对患者心理状态不佳的情况，平台应推出情感支持服务，在问诊界面提示医生关注患者心理问题，由医生在问诊过程中给予一定情感支持。四是针对患者疾病认知不足的情况，平台与医生应重视科普工作，对睡眠共病进行科普，尤其是针对难以自我判断、容易引起睡眠问题的其他疾病。医生在问诊过程中应注意患者睡眠共病问题，帮助患者找到影响睡眠的真正原因。

作者贡献：庞盼杏负责研究设计、数据收集与处理、论文撰写；何彩荣负责协助数据收集与处理；徐中岳负责提出研究选题、设计论文框架；张磊、陈景信、石荣丽、翁开源负责提供指导、论文审核。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 贾福军，李雪丽. 睡眠与睡眠障碍［J］. 中华全科医学杂志，2016，15（7）：497－499.
2 夏天吉，闫明珠，王智，等. 大小鼠失眠模型和评价方法研究进展［J］. 中国实验动物学报，2022，30（3）：428－435.
3 李佳琦，严悦蓉，余叶蓉. 睡眠障碍与糖尿病的关系及其对糖代谢的影响研究进展［J］. 中国全科医学，2017，20（11）：1300－1304.
4 杨璇，乔雨晨，赵洁，等. 认知障碍患者睡眠障碍评估工具的应用进展［J］. 中华护理杂志，2020，55（12）：1884－1889.
5 艾瑞咨询. 中国在线医疗健康服务消费白皮书［EB/OL］.［2023－09－05］. https://www.iresearch.com.cn/Detail/report?id=4057&isfree=0.
6 张东鑫，张敏. 图情领域 LDA 主题模型应用研究进展述评［J］. 图书情报知识，2022，39（6）：143－157.
7 陆泉，朱安琪，张霁月，等. 中文网络健康社区中的用户信息需求挖掘研究——以求医网肿瘤板块数据为例［J］. 数据分析与知识发现，2019，3（4）：22－32.
8 于本海，卢畅. 在线健康社区信息主题特征及其潜在价值研究——基于 LDA 模型对“百度痛风病吧”案例的分析［J］. 价格理论与实践，2022（3）：195－198，206.
9 余佳琪，赵豆豆，刘蕤. 在线健康社区慢性病患者评论主题情感协同挖掘研究——以甜蜜家园为例［J］. 数据分析与知识发现，2023，7（10）：95－108.
10 刘冰，历鑫，张赫钊，等. 网络健康社区中身份转换期女性信息需求主题特征及情感因素研究——以“妈妈网”中“备孕版块”为例［J］. 情报理论与实践，2019，42（5）：87－92.
11 CABRIELA G，KRISTINA V，CARINE K，et al. Women’s experience with stress urinary incontinence：insights from social media analytics［J］. The journal of urology，2019，203（5）：962－968.
12 LI Y，BRUCE R，THOMAS M A，et al. Leveraging latent Dirichlet allocation in processing free－text personal goals among patients undergoing bladder cancer surgery［J］. Quality of life research：an international journal of quality of life aspects of treatment，care and rehabilitation，2019，28（6）：1441－1455.
13 SOOWON P，YAEJI K K，Ah J S. Leveraging text mining approach to identify what people want to know about mental disorders from online inquiry platforms［J］. Frontiers in public health，2021（9）：759802.
14 STEVENS K，KEGELMEYER P，ANDRZEJEWSKI D，et al. Exploring topic coherence over many models and many topics［C］. Jeju Island：Conference on Empirical Methods in Natural Language Processing，2012.
15 张梅兰，侯晓聪. 社交媒体平台的癌症疾痛叙事与患者自我疗愈研究［J］. 新闻大学，2022（10）：101－117，123.
16 王萱，黄涛. 在线问诊患者特征和代问现象研究——以“丁香医生”为例［J］. 中国卫生政策研究，2021，14（9）：54－61.
17 饶诗彤，张可涵，曾燕，等. 生活习惯对老年人失眠的影响——基于多中心社区流行病学调查研究［J］. 现代预防医学，2023，50（4）：577－581，610.
18 胡世平，朱宏斌，王东旭. 功能性胃肠病与睡眠障碍的研究进展［J］. 医学研究生学报，2019，32（8）：861－865.
19 聂卉，吴晓燕，林芸. 基于在线问诊记录的抑郁症病患群组划分与特征分析［J］. 数据分析与知识发现，2022，6（Z1）：222－232.