基于 LDA 的罕见病患者健康需求研究——以血友病为例

梁娟芳 张军伟 师成虎 加妙丽 王雅婧 贺培风

（山西医科大学管理学院 太原 030001）

〔摘要〕 **目的/意义** 揭示以血友病为代表的罕见病患者群体潜在健康需求，为制定更精准有效的罕见病保障支持措施提供参考。**方法/过程** 以百度“血友病吧”帖子文本为数据来源，分析用户关注热点及情感倾向。**结果/结论** 获取有效帖子文本共 9 765 条，得出血友病患者关注的 7 大主题。网络用户对血友病相关问题的态度不容乐观。建议强化诊疗能力建设、构建多维度社会支持网络、持续完善医药保障体系、加强用药管理。

〔关键词〕 罕见病；血友病；健康需求；隐含狄利克雷分布

〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2025.10.006

Study on Health Needs of Patients with Rare Diseases Based on LDA: a Case Study of Hemophilia

LIANG Juanfang, ZHANG Junwei, SHI Chenghu, JIA Miaoli, WANG Yajing, HE Peifeng

School of Management, Shanxi Medical University, Taiyuan 030001, China

〔**Abstract**〕 **Purpose/Significance** To uncover the potential health needs of patient groups with rare diseases represented by hemophilia, and to provide references for formulating more precise and effective rare disease support and guarantee measures. **Method/Process** Using the post texts from Baidu's “Hemophilia Bar” as the data source, the hot topics of users' concern and their emotional tendencies are analyzed. **Result/Conclusion** A total of 9 765 valid post texts are obtained, and seven key themes of concern to hemophilia patients are identified. The attitudes of netizens towards hemophilia－related issues are not optimistic. It is recommended to strengthen the construction of diagnosis and treatment capabilities, establish a multi－dimensional social support network, continuously improve the medical insurance system, and enhance medication management.

〔**Keywords**〕 rare disease; hemophilia; health needs; latent Dirichlet allocation (LDA)

1 引言

血友病是罕见病领域的“大病种”，是凝血因子缺乏导致凝血功能障碍的遗传性疾病，患病率逐年攀升[1]。与大多数罕见病相似，血友病患者分布零散，其群体及家属的健康需求难以被社会广泛认

〔修回日期〕 2025－09－04

〔作者简介〕 梁娟芳，讲师，发表论文 10 余篇。

〔基金项目〕 国家社会科学基金项目（项目编号：21BTQ050）。

知和有效整合。

健康需求是个人或群体为达到、维持或改善健康状态，在生理、心理、社会方面的要求，包括疾病管理、心理社会、经济保障等需求。目前健康需求分析方法以社会调查访谈为主，存在样本量有限、主观偏差大等弊端。随着互联网普及，血友病患者及其家属将在线健康平台作为寻求支持、分享经验、交流信息的重要途径。此类平台中的海量文本，蕴含着血友病患者多方面的深层次需求与情感倾向。目前网络健康文本信息研究主要采用不同文本挖掘和分析技术，通过凝练知识主题及特征规律挖掘深层次需求[2]。万欣等[3]以百度“乙肝吧”为数据来源，采用 K－means 聚类算法与 CorEx－Topic 主题模型，解释传染病社区用户病耻感根源及情绪反应多样性。叶艳等[4]利用 Python 爬取“好大夫在线”高血压评论数据，借助 LDA－BiLSTM 模型，得到 6 大医疗服务质量主题并进行情感分析，剖析服务质量差的原因。周欢等[5]搜集在线健康社区评论数据，运用 TF－IDF、TextRank 和隐含狄利克雷分布（latent Dirichlet allocation，LDA）共 3 种方法揭示评论主题和情感分布特征，帮助用户和管理者识别虚假评论。

本研究借鉴上述研究思路，采用 LDA 主题模型与 SnowNLP 情感分析技术，以百度“血友病吧”用户帖子文本为数据源，剖析帖子文本中蕴含的复杂语义信息，揭示以血友病为代表的罕见病患者群体的潜在健康需求，为制定更精准有效的罕见病保障支持措施提供参考。

2 资料与方法

2.1 数据获取与预处理

登录百度“血友病吧”，爬取 2011 年 1 月 1 日—2025 年 5 月 13 日用户发表的所有主帖，共计 26 899 条。经过去重、删除无效帖子、填充缺失值、剔除小于 5 字的短文本，得到有效帖子文本 9 765 条。根据帖子内容构建“血友病用语”自定义词典；在哈尔滨工业大学停用词表基础上，结合帖子内容添加无意义词至停用词表。采用 Python 的 jieba 库对文本进行分词。

2.2 研究方法

2.2.1 LDA 主题模型分析 LDA 主题模型可将非结构化文本数据转化为具备内在规律的结构化数据，在此基础上挖掘文档中隐藏的核心主题，并分析主题的具体内容构成及其动态变化趋势[6]。采用困惑度指标确定文档主题数量，再用 pyLDAvis 软件对抽取主题动态可视化。

2.2.2 SnowNLP 情感分析 情感分析是对文本蕴含的情感信息进行分析、处理、归纳和推理的过程，可了解网络用户隐藏的情感态度、观点和情绪[7]。SnowNLP 是用于中文文本情感分析的 Python 库，自带中文正负情感训练集，能根据输入文本特征计算情感值。利用 SnowNLP 计算 9 765 条帖子文本的情感值。情感值区间为［0，1］，设定［0，0.4］为消极情感，［0.4，0.6］为中性情感，［0.6，1］为积极情感[7]。

3 结果

3.1 主题分析结果

3.1.1 LDA 主题提取 利用 Python 调用 Gensim 软件包，超参数设为默认值，进行 LDA 主题模型分析。主题数量－困惑度折线图，见图 1。主题数量为 7 时模型效果较好，主题间关系分明，因此将主题数量确定为 7。

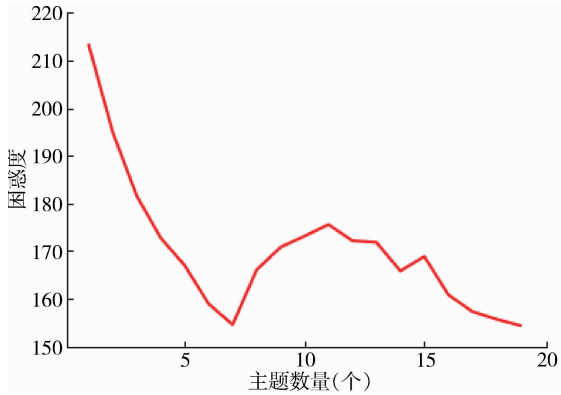

图 1 主题数量－困惑度折线

血友病用户帖子文本的 LDA 主题可视化，见图 2。左侧每个圆圈代表不同主题，圆圈大小代表主

题所占比重[6]。对评论文本进行 LDA 主题分析后得到主题－词汇矩阵和文档－主题矩阵。根据主题－词汇矩阵提取 7 个主题及前 10 个主题关键词，各主题占比相当，见表 1。

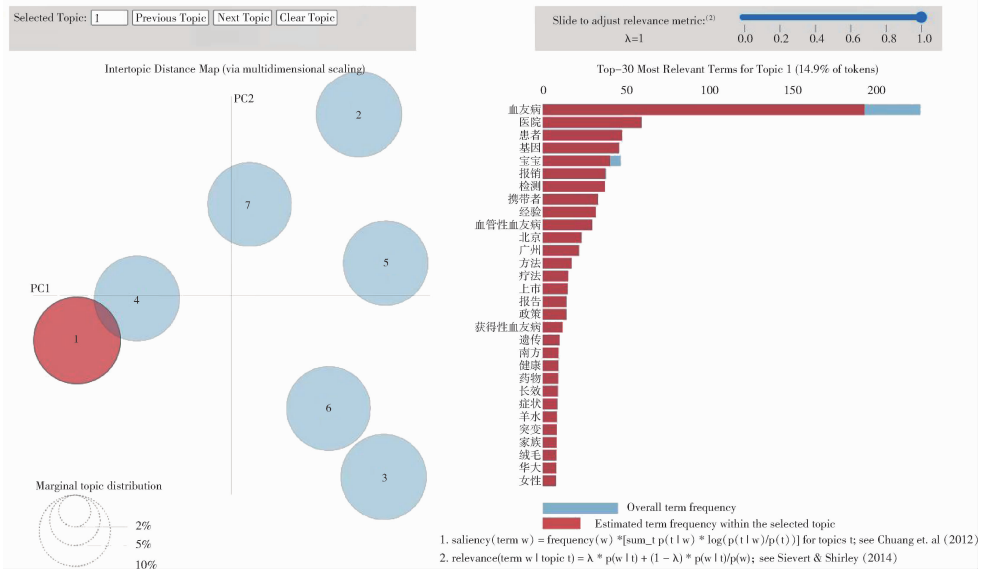

图 2 血友病用户贴子文本的 LDA 主题可视化

表 1 基于 LDA 主题模型的血友病用户贴子文本主题内容

序号	主题名称	前 10 个主题关键词	占比（%）
1	基因遗传	血友病、医院、患者、基因、宝宝、报销、检测、携带者、经验、血管性血友病	14.9
2	护理处置	凝血因子、注射、止血、护理、经验、拔牙、输液、疫苗、小孩、处置	14.5
3	症状表现	出血、关节、膝盖、淤青、尿血、牙龈、脚踝、肿胀、疼痛、伸不直	14.4
4	预防治疗	因子、治疗、膝关节、置换、手术、预防、百因止、髋关节、效果、血友病	14.3
5	医保报销	医院、医保报销、援助、比例、申报、慢性病、异地、门诊、负担、沉重	14.1
6	药物可及	买药、甲型血友病、当地、买不到、县城、断药、进药、国产、拿药、疗效	13.9
7	生活就业	生活、工作、入职、不敢、隐瞒、甲型血友病、大学、公务员、歧视、放弃	13.9

3.1.2 主题内容分析 根据文档－主题概率分布矩阵，选取每个帖子文本主题概率分布最大值对应的主题作为该文档所属主题，此文本即为主题支持文档。参照各主题关键词及对应文档，归纳 7 个主题含义。主题 1：基因遗传。聚焦血友病患者及基因携带者降低该病遗传给后代风险的方法。当前可通过基因检测技术，在胎儿出生前识别其是否携带特定罕见病基因。主题 2：护理处置。主要涉及血友病患者面对疫苗接种、拔牙或输液等操作的护理经验。由于体内凝血因子不足，上述操作后可能出血不止，必要时须提前注射凝血因子。主题 3：症状表现。血友病患者常出现皮肤淤青、牙龈出血、鼻出血、尿血等自发性出血症状。长期反复出血会导致关节病变或残疾，多数成人患者伴有关节肿胀、疼痛、伸不直现象。主题 4：预防治疗。主要关于血友病的预防和治疗，目前通过凝血因子替代进行预防和按需治疗。预防治疗是在无症状时规律补充凝血因子，按需治疗是出现症状后补充凝血因子止血，常伴并发症须治疗。主题 5：医保报销。主要内容是不同地域与政策下，血友病患者医保报销比例不同。患者需长期药物治疗，虽有医保和企业援助，但治疗直接和间接成本依然很高。主题 6：药物可及。讨论新药陆续上市但全国普及程度不一的问题。一些医院血友病治疗药品供应不连续，时有断药情况，偏远地区患者拿药更困难。主题 7：生活就业。聚焦血友病患者在日常生活和职场就业

中遭遇的困境与歧视。病情有效控制的患者可正常生活，病情控制差的患者可能面临工作障碍。

3.1.3　主题演化分析　计算各主题在当年总体研究中的相对重要性，即各主题每年的主题强度，通过主题强度变化趋势分析血友病患者及家属健康需求的演变特征，见图 3。基因遗传主题强度 2015—2016 年明显增强；医保报销主题 2018—2019 年关注度最高；护理处置、症状表现、预防治疗 3 个主题强度 2019—2022 年显著提升。原因可能为：技术迭代、市场竞争、规模效应和政策支持推动下，2015 年基因检测价格大幅下降，血友病患者对基因检测关注度走高；2018 年 5 月国家卫生健康委员会将血友病纳入《第一批罕见病目录》，释放医保倾斜信号；2019 年国家医保谈判大幅降低重组凝血因子价格，部分省市试行罕见病专项基金或大病保险；2020—2022 年突发公共卫生事件期间，医院就诊受限，患者更依赖家庭护理和自我管理，自我注射技巧、出血判断等信息需求激增。

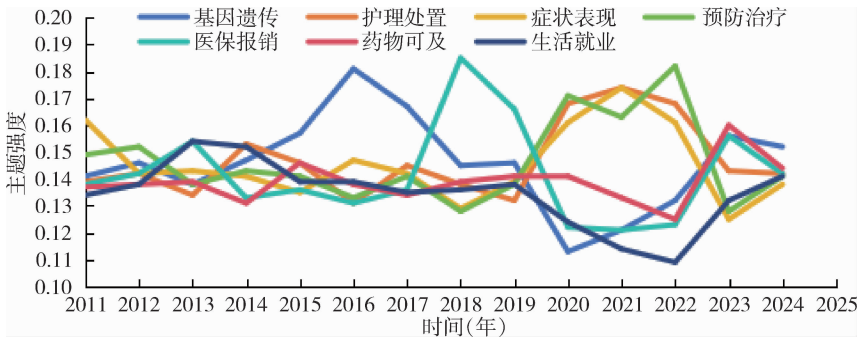

图 3　血友病用户帖子文本主题强度演化

3.2　情感分析

对文本数据进行情感分析，获积极情感文本 2 640 条，消极情感文本 5 039 条，中性情感文本 2 086 条。总体来看，积极情感占比 21% ~31%，消极情感占比 48% ~57%，中性情感占比 17% ~25%，见图 4。各主题消极情感占比均高于积极和中性情感，网络用户对血友病相关问题态度不容乐观。

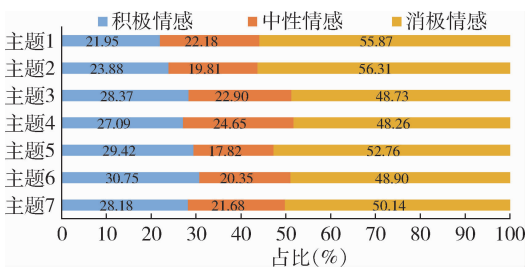

图 4　各主题不同类别情感占比

根据分词结果，取频率最高的前 100 个词制作情感词云图，见图 5。词出现频率越高，字越大，离中心越近。前 20 个高频词及相对词频，见表 2。结合词云图与高频词表，“血友病”和“因子”词频远高于其他关键词，表明血友病和因子是最受关注的对象，也是影响患者及家属情感的重要载体。出血、关节问题、基因检测、医保报销、手术治疗是患者及家属情感评价的主要内容。

图 5　整体情感词云图

表 2 前 20 个高频词及相对词频

序号	关键词	词频（%）
1	血友病	3.17
2	因子	2.13
3	医院	0.91
4	出血	0.81
5	关节	0.61
6	病友	0.60
7	治疗	0.59
8	孩子	0.56
9	基因	0.54
10	报销	0.45
11	宝宝	0.43
12	患者	0.42
13	手术	0.41
14	抗体	0.40
15	膝关节	0.38
16	检测	0.35
17	置换	0.34
18	携带	0.33
19	检查	0.31
20	预防	0.29

4 讨论

4.1 医疗防治

按血友病用户健康需求内容，将 7 个主题归纳为 4 大类。医疗防治大类包含主题 1“基因遗传”、主题 2“护理处置”、主题 4“预防治疗”，共占比 43.7%，是患者及家属最关注的主题。基因遗传是血友病根源，基因检测虽能评估风险，但高遗传风险仍使患者家庭焦虑。血友病患者接受预防性治疗后可维持正常生活、工作和学习，但防治面临多重挑战，如拔牙、输液等业务需特殊护理处置。此类主题中，积极情感文本占比 24.27%，血友病虽是终身伴随疾病，但积极治疗和预防，尤其是对儿童患者的预防性治疗，能保证患儿正常生活质量，患儿心态整体相对平和。消极情感文本占比 53.53%，基层医疗体系中，许多医生对血友病了解有限，误诊、误治现象频发，甚至延误病情。

4.2 公共支持

该主题包括主题 3“症状表现”和主题 7“生活就业”，共占比 28.3%。血友病是需终身用药的罕见病，症状多样且严重，从轻微淤青到危及生命的出血，患者常伴身体不便及残疾，公共卫生服务机构应加强知识普及，引导患者理性认知疾病，构建包容性就业环境，减轻患者心理负担。该主题积极情感评论文本占比 28.28%，残疾人联合会、社工、志愿者等社会组织提供的有效支持，对血友病患者生活就业帮助很大。消极情感评论文本占比 49.42%，血友病群体身体特殊性给生活和工作带来不便，影响生活质量，引发自卑与病耻感。

4.3 医保报销

该主题占比 14.1%。多数成人血友病患者有关节严重变形并发症，需理疗、康复甚至关节置换，医保报销比例各地差异大。虽血友病已纳入《第一批罕见病目录》，但诊疗支持、报销比例及社会支持仍不足。该主题支持文档中，积极情感评论文本占比 29.42%，国家近年来对罕见病群体关注度提升，制定相应医疗保障政策，对此患者群体较满意。消极情感评论文本占比 52.76%，该病间接成本高且须长期用药，医保报销不能大幅缓解患者家庭经济负担。

4.4 药物可及

该主题占比 13.9%。长期注射凝血因子是血友病预防治疗的关键，但药物可及性问题显著影响患者治疗体验。患者普遍反映县级及社区医院药物短缺，要到省级三甲医院购药。同时药物品牌、剂量及患者个体差异影响疗效，保障优质药品供应稳定与经济性至关重要。该主题支持文档中，积极情感评论文本占比 30.75%，陆续上市的新药及政策支持可使患者获得更好的治疗效果。消极情感评论文本占比 48.90%，该病属罕见病，许多基层医院缺乏治疗药物，药物供应不及时，缺医少药加剧患者困境。

5 建议

5.1 强化诊疗能力建设，使罕见病患者便捷就医

一是降低罕见病遗传概率。通过家族基因检测、婚检、试管婴儿、产检等技术预防出生缺陷，实现优生优育。二是加强基层医务人员罕见病诊疗培训。将服务下沉，确保患者在基层医疗机构得到有效指导和治疗，满足便捷就医需求，减轻负担。三是建立完善的罕见病诊疗协作机制[8]。多数罕见病患者须终身服药及定期评估机体组织受损程度，可在基层医院完成，同时进行双向转诊，提升综合协同服务能力。

5.2 建立全方位社会支持网络，使罕见病患者不再孤单

一是搭建罕见病患者交流平台。在线健康社区可聚集不同地区患者，方便同类罕见病群体学习交流，提升归属感。二是为罕见病患者提供更多就业机会。拓展就业渠道，帮助其实现自我价值。三是进行罕见病宣传教育，提高公众认知，消除歧视偏见。

5.3 持续完善医药保障体系，降低罕见病患者经济负担

一是进一步扩大罕见病药物种类和并发症医治覆盖范围，提高报销比例，降低自付比例。二是建立罕见病多层次医疗保障制度，罕见病用药多为高值药物，单一医保制度难以承担，建议财政支持、大病医保、医疗救助、商业保险、慈善保障等多方参与[9]。三是出台罕见病专项保障政策。参照先进地区经验，创新保障机制[1]。

5.4 加强用药管理，使罕见病患者有药可用

一是加大罕见病领域创新药物研发力度，优化药物准入流程，支持新治疗方式及产品。二是通过建立不同罕见病病种患者信息数据库，结合生物信息学、基因组与蛋白质组学、物联网医学等新技术，开发更多惠及患者的治疗药物或器械[6]。三是扩大罕见病定点医院数目，督促医院及时提供药品，确保患者有药可用。

6 结语

与传统研究依赖调查访谈不同，本研究采用 LDA 主题模型与情感分析方法，基于网络文本挖掘血友病患者健康需求，并以血友病为罕见病代表提出建议。未来可结合多模态数据，深入分析患者行为轨迹，构建动态需求演变模型，提升健康服务精准性与实时性。

作者贡献：梁娟芳负责数据分析、论文撰写与修订；张军伟、师成虎负责数据分析、论文撰写；加妙丽、王雅婧负责数据分析；贺培风负责提供指导。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 杨仁池．《中国血友病管理指南》（2024 版）［M］．北京：中国协和医科大学出版社，2024.

2 RANARD B L，WERNER R M，ANTANAVICIUS T，et al. Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care［J］. Health affairs，2016，35（4）：697－705.

3 万欣，袁海曼．在线健康社区用户情绪识别与分析［J］．医学信息学杂志，2023，44（12）：15－19.

4 叶艳，吴鹏，周知，等．基于 LDA－BiLSTM 模型的在线医疗服务质量识别研究［J］．情报理论与实践，2022，45（8）：178－183.

5 周欢，刘嘉，张培颖，等．复杂网络视角下在线健康社区评论有用性研究［J］．情报科学，2022，40（9）：88－97.

6 邰杨芳，郭樱，王紫琼，等．基于公众关注主题及情感视角的我国罕见病医药保障策略研究——以诺西那生钠为例［J］．中国药房，2023，34（7）：774－779.

7 张冬，魏俊斌．情感驱动下主流媒体疫情信息数据分析与话语引导策略［J］．图书情报工作，2021，65（14）：101－108.

8 张永理，张开然．我国罕见病患者医疗保障现状及改进建议［J］．卫生经济研究，2023，40（1）：53－56.

9 李常印，殷玉茹．罕见病医疗保障机制研究——以血友病为例［J］．中国医疗保险，2023（7）：104－110.