

• 健康教育 •

基于 LDA 主题模型的网络问答社区妊娠合并糖尿病孕产妇健康信息需求分析

马誉萁¹, 侯小妮¹, 段红梅¹, 徐杨², 陈延亭², 李博文¹, 金宁宁¹

摘要:目的 应用 LDA 主题模型挖掘妊娠合并糖尿病孕产妇信息需求,为医护人员开展有效信息服务提供参考。方法 采用网络爬虫程序获取网络问答社区妊娠合并糖尿病板块的用户问答数据,采用 LDA 主题模型对数据进行主题识别,分析信息需求。结果 通过对 126 616 条数据的 LDA 主题分析,妊娠合并糖尿病孕产妇的信息需求包括 8 类:母儿危害与检查,母儿自我监护,营养体重质量控制与胎儿生长发育,筛查与产检,病友经验分享与情感支持,细分食物类别的选择与升糖效应,分娩及产后母婴健康管理,血糖调控与餐食、胰岛素用药。结论 妊娠合并糖尿病孕产妇的信息需求多元、专业、精细化,包括认知、生理、情感等多个维度,医护人员应发挥医患共同信息服务主体作用,针对妊娠合并糖尿病孕产妇提供专业、精细、全面的信息服务。
关键词:妊娠合并糖尿病; 健康信息需求; 网络问答社区; LDA 主题模型; 孕期保健; 围生期护理
中图分类号:R473.71;R473.2 **DOI:**10.3870/j.issn.1001-4152.2023.07.086

Analysis on health information needs of women with gestational diabetes in online question and answer communities based on LDA theme model Ma Yuqi, Hou Xiaoni, Duan Hongmei, Xu Yang, Chen Yanting, Li Bowen, Jin Ningning. School of Nursing, Beijing University of Traditional Chinese Medicine, Beijing 100102, China
Abstract: **Objective** To excavate the information needs of women with gestational diabetes based on latent Dirichlet allocation (LDA) theme model, and to provide reference for medical workers to carry out effective information services. **Methods** The users' question and answer (Q&A) data of gestational diabetes in online Q&A communities were obtained by using the web crawler program. The LDA theme model was used to identify the theme of the data and analyze the information needs. **Results** Through LDA thematic analysis of 126,616 pieces of data, the information needs regarding gestational diabetes were factored into eight categories: maternal and infant harm and inspection, maternal and infant self-monitoring, nutritional weight control and fetal growth and development, screening and prenatal examination, experience sharing and emotional support of patients, the selection of food categories and sugar raising effect, maternal and infant health management during and after childbirth, blood glucose control and diet, and insulin medication. **Conclusion** The information needs of women with gestational diabetes are diversified, professional and meticulous, with the needs falling into multiple dimensions of cognition, physiology and emotion. Medical workers should play their role as the mainstay in information service, and provide professional, sophisticated and comprehensive information services for gestational diabetes.
Key words: gestational diabetes; health information needs; online Q&A community; latent Dirichlet allocation model; pregnancy care; perinatal care

妊娠合并糖尿病包括孕前糖尿病和妊娠期糖尿病两种类型,孕前糖尿病是在原有糖尿病基础上合并妊娠;妊娠期糖尿病指妊娠前糖代谢正常,妊娠期诊断的糖尿病^[1]。2021 年国际糖尿病联盟发布的数据显示,我国 20~49 岁育龄期妇女中妊娠合并糖尿病患病率为 8.6%^[2],并逐年递增。妊娠合并糖尿病病情复杂多变,可对母儿产生严重危害。在妊娠、分娩和产后不同阶段,妊娠合并糖尿病孕产妇既是自身健康照顾者,更是胎儿、新生儿健康主要照顾者,为尽可能保证母婴健康并解决和应对围生期复杂问题,妊娠合并糖尿病孕产妇会因信息不足产生多维度的健康

信息需求。信息需求是激发信息获取行为产生的原动力,多项研究显示,互联网已成为孕产妇获取健康信息的主要渠道^[3-5]。人们可以通过搜索引擎链接到某个网站获取健康信息,也可以通过社交媒体进行信息共享交流来满足信息需求,获得支持与应对方法。因此,对网络数据进行分析与挖掘是了解用户健康信息需求的重要途径。

LDA(Latent Dirichlet Allocation)主题模型^[6]是语料库生成概率模型,可对数据进行主题聚类或文本分类,依据文本中词与词的共现概率挖掘文本中隐含的主题信息,广泛应用于文本特征挖掘等领域。目前越来越多研究者通过 LDA 主题模型分析互联网用户的健康信息需求,刘浩等^[7]、安欣宇等^[8]、仁青卓玛等^[9]均通过 LDA 方法对网络问答社区数据进行挖掘,提炼出了哮喘儿童、孕产妇等领域的信息需求。本研究通过 LDA 主题模型对国内常用网络社区妊娠合并糖尿病相关用户评论、留言、问答数据进行挖

作者单位:1. 北京中医药大学护理学院(北京,100102);2. 中日友好医院产科
马誉萁:女,硕士在读,学生
通信作者:金宁宁,ningning_jin@163.com
科研项目:中华医学会杂志社 2021-2022 年护理学科研究课题(CMAPH-NRI2021062)
收稿:2022-11-07;修回:2022-12-28

掘,了解妊娠合并糖尿病孕产妇用户的健康信息需求,为医护人员建设和完善针对性信息服务提供参考。

1 资料与方法

1.1 网络数据来源及采集 妊娠合并糖尿病为专科常见疾病,因此首先通过百度搜索引擎以“妊娠合并糖尿病”“糖尿病”“糖妈妈”等为关键词搜索得到包括“糖友网”“糖尿病网”“妈妈网”“丁香妈妈”和“妈妈育儿网”5 个相关健康网站。浏览每个网站详情页后发现“糖尿病网”“妈妈网”2 个网站的互动问答数据丰富,因此选择其作为健康类网络问答社区代表。综合问答社区也是人们交流健康话题的重要平台,本研究选择“知乎”为知识问答社区代表,选择“抖音”为热门短视频社区代表,选择“新浪微博”为基于用户关系的社交媒体代表,选择“百度贴吧”为社区论坛代表。共选取上述 6 个网络问答社区提取问答信息,使用八爪鱼和后羿采集器采集自平台建立以来至 2022 年 5 月 1 日的所有问答,包括发帖标题、帖子详情和评论详情共 256 600 条文本数据。对原始数据使用 Excel 版本软件去除重复项处理,并将去重后的数据进行降序排列,删除排序后数据末尾的大量语气助词、连接词、数字、英文字母、标点符号、不完整语句和不包含信息需求内容的文本单元格,最后共留有 126 616 条有效数据。

1.2 方法

1.2.1 数据预处理 使用 Python 中的 jieba 分词库进行分词。为弥补 jieba 包对专业词汇的分词处理缺陷,本研究以《医学词汇大全》^[10] 和《妊娠合并糖尿病——临床实践指南》^[11] 作为自定义词典供机器学习,参考中文停用词表和哈尔滨工业大学停用词表^[12] 将得到的分词结果中包含的大量语气助词、连接词、数字、英文字母、标点符号等停用词去除。此外,考虑用户自然语言交流的不规范性,对常见词汇如“怀孕”和“妊娠”“糖化”和“糖化血红蛋白”等进行同义词转化处理。

1.2.2 LDA 模型分析 LDA 主题模型可将文档集中的每篇文档主题以概率分布形式给出,用于推测文档的主题分布,进行主题聚类或文本分类^[6]。首先使用 Python 对已得到的高频词进行 *k*-means 聚类分析,计算聚类结果为 2~29 的困惑度,困惑度越低表示聚类效果越好^[13],确定聚类结果为 10。最后利用轮廓系数法对聚类结果进行评估,轮廓系数值越大表示聚类效果越好,确定主题数为 10,以此为依据识别归纳主题类别。调用 Python 中的 gensim 库建立 LDA 主题模型,利用非监督机器学习来挖掘大文档的潜在主题,将聚类主题数 *k* 设置为 10 运行结果得到 10 个主题及每个主题下的词,分析 9、10 两个主题下的关键词后无法得出有效主题,最终将主题数调整为 8 个。

2 结果

2.1 数据一般特征 本研究共选取 6 个网络问答平台的妊娠合并糖尿病问答数据共 126 616 条,采集数据的基本情况如表 1 所示。Python 分词后的到高频词统计,其中前 10 项分别是血糖值、饮食、孩子、医生、控制、孕产妇、胰岛素、空腹、小时、餐后血糖。应用微词云对频数大于 100 的高频词绘制词云图,由图 1 可见,血糖值、饮食、胰岛素、控制、孩子等是妊娠合并糖尿病用户重点关注的内容;检查、糖耐量试验、空腹、餐后血糖、影响等方面为次要关注内容。

表 1 妊娠合并糖尿病互联网平台问答数据采集基本情况

网站名称	类别	网址	数量(个)
糖尿病网	糖尿病健康社区	https://www.tnbz.com	97994
百度贴吧	社区论坛	https://www.baidu.com	13207
妈妈网	母婴健康社区	https://www.mama.cn	10596
新浪微博	社交媒体	https://www.weibo.com	1731
知乎	知识问答社区	https://www.zhihu.com	1599
抖音	短视频社区	https://www.douyin.com	1487



图 1 妊娠合并糖尿病高频词词云

2.2 问答数据的主题特征 LDA 主题模型将 126 616 条规范问答数据划分为 8 个主题,每个主题及前 20 位相应关键词如表 2 所示。具体为:①母儿危害与检查。该主题反映用户十分重视妊娠合并糖尿病对母亲、胎儿和新生儿的危害及相关检查与结果判断信息,其中对疾病危害的关注主要集中在母亲羊水过多,胎儿巨大儿和胎儿畸形等方面。②母儿自我监护。该主题提示用户需要自身与胎儿健康监护相关的知识信息,其中对自身的监护主要为血糖监测,对胎儿的监护包括胎心和胎动等内容。③营养体质量控制与胎儿生长发育。该主题反映用户希望知道如何做才能达到营养摄入、体质量控制和胎儿正常生长发育的平衡。④筛查与产检。该主题体现用户需要妊娠合并糖尿病相关的筛查与产检知识与信息,其中筛查和产检都包括具体项目、费用、注意事项、必要性和结果解读等认知和决策信息需求。⑤病友经验分享与情感支持。该主题提示孕产妇渴望进行病友交流,包括经验分享、倾诉心事、互相鼓励、互相监督等,同时患病母亲心理压力巨大,家庭支持也不可或缺。⑥细分食物类别的选择与升糖效应。该主题反映患者选择具体食物时要考虑食物的升糖效应而存

在选择困难,孕产妇需要食物的分类、参数、用量、搭配、烹饪、品牌等具体食物选择信息。⑦分娩及产后母婴健康管理。该主题体现患者需要分娩知识如分娩方式、时间选择以及住院催产等相关信息;产后母

婴健康管理包括疾病预后、新生儿健康、喂养方式等方面。⑧血糖调控与餐食、胰岛素用药。包括不同时间的血糖值调控范围,血糖调控与饮食、活动及胰岛素使用的具体关系,以及低血糖的识别、预防与应对。

表 2 LDA 聚类形成的 8 个主题、前 20 位关键词及问答原文范例

主题	关键词	问答原文范例
① 母儿危害与检查	影响,诊断,发生,新生儿,胎儿,风险,并发症,筛查,治疗,检查,合并,异常,眼底,危险,症状,巨大儿,高危,畸形,血压值,高血压	①胎儿偏大会不会是巨大儿? ②感觉视物模糊,检查说是黄斑水肿眼底出血,怎么办呀? ③空腹血糖 12 点多,彩超显示胎儿泌尿系畸形,怎么办?
② 母儿自我监护	血糖高,第二胎,努力,胎心,超过,辛苦,胎动,第一胎,预产期,孕晚期,孕早期,方案,休息,孕中期,痛苦,子宫收缩,保胎,照顾,小心,监护	①怎样自己监测血糖? ②今天去做胎心监护,两次都没合格,为什么? ③胎动怎么数呀?
③ 营养体质量控制与胎儿生长发育	胎儿,增加,体重,营养,发育,准妈妈,食物,摄入,胎盘,减少,运动,维生素,脂肪,热量,蛋白质,补充,影响,能量,产妇,吸收	①营养能保证吗? 我 3 个月瘦了 10 斤,②每天都是半饥饿状态,真不知道宝宝营养够不够!
④ 筛查与产检	医院,检查,糖耐量试验,住院,建议,二甲双胍,复查,检测,产检,喝糖水,估计,结果,1 个月,顺利,B 超检查,咨询,抽血,甲胎蛋白,糖人,内分泌科	①测糖耐不喝糖水只吃早餐去检查,有影响吗? ②一下抽了 3 次血,有 1 次数值不达标就判定我为妊娠期糖尿病,不合理吧? ③产检做无创 DNA 还是羊水穿刺?
⑤ 病友经验分享与情感支持	孩子,感觉,希望,打针,肚子,血糖仪,心情,吃药,宝妈,经验,身体,平安,纠结,朋友,坚强,问问,没什么大不了,糖尿病病友,爱吃糖,几个月	①妊娠期糖尿病,有没有有经验的宝妈? ②每天都要自己打针煮饭还得上班,真苦!
⑥ 细分食物类别的选择与升糖效应	水果,主食,不吃,喜欢,牛奶,米饭,甜食,鸡蛋,蔬菜,粗粮,含糖量,餐后 1 小时血糖,馒头,苹果,血糖生成指数,加餐,面包,饼干,黄瓜,无糖	①网上卖的那些无糖糖果、饼干、蛋糕能吃吗? ②你们主食加餐都吃什么? 肚子饿血糖又要正常怎么吃?
⑦ 分娩及产后母婴健康管理	孩子,健康,担心,次数,母乳喂养,餐后,小时血糖,1 型糖尿病,遗传,顺产,产妇,出生,幸福,奶粉,2 型糖尿病,剖腹产,羊水,放心,生病,麻烦	①孕 38 周顺产好还是剖腹产好? ②今天 40 周了,宝宝还没要出生的意思,是否要催产? ③血糖控制到多少能母乳喂养? ④产后还需要控制饮食吗?
⑧ 血糖调控与餐食、胰岛素用药	空腹,小时,餐后血糖,血糖值,低血糖,晚餐,早餐,睡前,加餐,胰岛素,午餐,餐前,酮体,诺和灵,诺和锐,长期有效,带胰岛素泵,增加数量,散步,剂量	①早餐真不知道吃什么了,吃点血糖就高! ②餐前 5.0 mmol/L,餐后 2 h 7.6 mmol/L 是不是有点高? 标准值是多少? ③你们都打几个单位的胰岛素,注射胰岛素的剂量这样多么?

3 讨论

3.1 对网络问答记录的 LDA 主题挖掘分析有助于发现妊娠合并糖尿病用户的隐性信息需求 LDA 主题模型通过建模来挖掘文本中的隐含主题,最大程度地保留表示文本含义的词语,通过衡量文档之间的语义相似性聚类主题,准确挖掘文本的隐藏信息,达到提取期望信息的目的。该研究方法在网络问答社区的文本挖掘中占独特优势,对挖掘互联网用户信息需求有重要意义。LDA 模型挖掘妊娠合并糖尿病孕产妇在线需求,能够发现国内大多数医院对患者健康教育的不足之处,也能丰富访谈、问卷等方式对患者信息需求的研究。

Taylor^[14]将信息需求划分为显性信息需求和隐性信息需求。目前对妊娠合并糖尿病孕产妇信息需求的研究常采用自制调查问卷^[15-16]或访谈法^[17-18],挖掘的主要是认知方面的显性需求信息,对隐性需求信息的挖掘不足。网络互动问答社区的高互动性以及隐私保护性,使得网络问答社区成为用户表达隐性需求信息的重要信息场所^[19-20]。检索文献发现医院对妊娠合并糖尿病孕产妇的常规健康宣教^[21-22]基本包括疾病一般知识、饮食原则、运动原则、血糖监测、药物治疗原则、并发症识别等方面。从本研究对互动问答社区信息需求的挖掘中可见,患者具有很多精细化

的信息需求,如食物的具体选择,从何时吃、吃什么、食谱、搭配、食物类别,到食物品牌、食物替代品,以及怎么吃、如何烹调等;患者还需要检查治疗决策相关的信息,如筛查方式选择、分娩方式选择、母儿并发症的治疗决策信息等;此外,疾病管理及病情变化的特殊情境常常使他们无助和焦虑,他们需要病友分享经验及情感支持,这些都属于用户隐性需求信息的范畴。用户在论坛中发布目前面临的问题以寻求解答,也将平台作为表达情绪和分享生活、经验的空間,用文字记录下患病全程遇到的各种困难,因此网络数据能反映用户的隐性需求信息,这些文本经过计算机处理便能得到反映妊娠合并糖尿病用户显性和隐性需求信息的关键主题。挖掘妊娠合并糖尿病孕产妇的隐性需求信息有助于医护人员丰富和补充常规健康宣教内容,为患者提供更为全面的信息服务。

3.2 妊娠合并糖尿病母亲存在多元化、精细化、专业性的信息需求 本研究得到的 8 个信息需求主题基本涵盖妊娠合并糖尿病孕产妇在妊娠、分娩及产后不同阶段的信息需求。Wilson^[23]提出信息需求是为解决情境中的问题,而产生与生理、认知和情感需求等满足相关的,以信息为主要目的的需求。本研究 LDA 主题挖掘的结果与 Wilson^[23]的信息需求理论契合,妊娠合并糖尿病孕产妇的信息需求包括生理、

认知和情感 3 个维度。认知相关信息需求主要为母儿危害与检查,新确诊的孕产妇疾病知识缺乏,此刻她们最想了解的信息是疾病对自身和胎儿的危害,部分孕产妇还会对疾病诊断标准的合理性持疑,此外还希望了解产检、特殊检查项目等注意事项和如何查看检查结果单等专业知识。医护人员需尽早进行疾病知识和相关检查知识的健康教育,减轻孕产妇确诊时紧张、焦虑和恐慌等不良情绪,同时教授孕产妇母儿自我监护技能,便于异常情况的识别与应对,从而减少并发症和不良妊娠结局的发生。临近分娩和产后的母亲还会有关于分娩方式、产后自身和胎儿健康管理的专业知识需求,因此医护人员应为不同阶段的孕产妇进行针对性健康教育,同时开展在线咨询,便于患者随时随地获取可靠、权威、准确的信息。生理相关信息需求主要是非药物治疗,即饮食、运动、体质量控制、血糖监测和药物治疗,从主题③、⑥、⑧可以看出孕产妇的饮食需求十分精细,提示医护人员应重视信息服务内容的细化,提供详细且实用的控糖信息,建立系统化、个性化的控糖方案;还可以进行目标管理^[24],构建在线控糖管理平台,提供在线饮食、运动、体质量、血糖记录功能,通过各项数值的监测和反馈来调整控糖方案,提高患者自我管理依从性^[25],达到有效控制血糖的目的。情感相关的信息需求主要是病友经验分享与情感支持,从主题②中的辛苦、痛苦,主题⑤中的纠结、坚强,以及主题⑦中的担心、麻烦等关键词可以看出,妊娠合并糖尿病孕产妇在确诊后有对自身和孩子健康的担心以及控糖压力,她们渴望通过与病友交流来得到鼓励支持与经验分享,网络问答社区不仅是疾病与健康信息获取和交流的平台,更是用户获得情感交流与支持的重要途径^[19-20],而互联网信息场是陌生人间的信息互动,较好的隐私保护性使得网络互动有利于个体情感情绪释放与交流,从同伴处获得同理和共情更是她们满足情感需求的重要方式,因此医护人员可以大力发展线上互动,发挥专业引导性,开展线上课程,提高患者知识水平并促进患者交流,改善患者负性情绪^[26],也可以将患者纳入信息服务提供者中,发挥医患结合的信息服务主体作用。

4 小结

本研究基于 LDA 主题模型对常用网络问答社区妊娠合并糖尿病相关问答记录进行数据挖掘与分析,提炼出 8 类主题,涵盖认知、生理、情感需求等多个维度,具有多元化、专业性、精细化等特征。医护人员应从妊娠合并糖尿病患者的需求出发,发挥医患共同信息服务主体作用,以提供更专业、精细、全面的信息服 务。妊娠合并糖尿病用户在网络问答社区发表的信息需求较为碎片化,问答数据提供的信息需求在系统性和情境性方面存在不足,未来需进一步结合定性访谈等方法获得更为详细的情境信息,以便为患者提供

更具有针对性和区分性的信息服务。

参考文献:

[1] American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes-2021[J]. Diabetes Care, 2021, 44(1): 15-33.

[2] International Diabetes Federation. IDF diabetes atlas 10th edition 2021 [R/OL]. (2021-11-02) [2022-09-30]. <https://diabetesatlas.org/data/en/country/42/cn.html>.

[3] 章娟,张志力,王倩,等. 孕产妇母婴健康信息辨别能力及其影响因素研究[J]. 中国全科医学, 2021, 24(24): 3065-3070.

[4] 张志力,李汝德,程晓芸,等. 孕产妇健康信息交流的人际网络研究[J]. 中国全科医学, 2020, 23(11): 1399-1405.

[5] 李星颖,陈维怡,金龙妹,等. 新冠疫情下从官方新媒体途径获取孕产保健信息与防护知识行为的横断面调查[J]. 复旦学报(医学版), 2021, 48(4): 520-526.

[6] Blei D M, Ng A, Jordan M I. Latent dirichlet allocation [J]. JMLR, 2003, 3(4-5): 993-1022.

[7] 安欣宇,陈育新,张晗. 网络健康社区用户的新冠肺炎信息需求[J]. 中华医学图书情报杂志, 2021, 30(2): 53-58.

[8] 王艺锦,何旭东,段红梅. 在线医疗社区中银屑病患者健康信息需求的分析[J]. 中华现代护理杂志, 2022, 28(16): 2131-2136.

[9] 仁青卓玛,王世奇,侯小妮,等. 在线母婴论坛中不同时期孕产妇健康信息需求的分析[J]. 中华现代护理杂志, 2022, 28(16): 2121-2125.

[10] 爱问文库. 医学词汇大全[S/OL]. (2021-02-26) [2022-09-30]. <https://ishare.iask.sina.com.cn/f/1c40Z5MtrT7.html>.

[11] 杨慧霞. 妊娠合并糖尿病——临床实践指南[M]. 北京: 人民卫生出版社, 2008: 9-11.

[12] Github. Stopwords[S/OL]. (2018-11-01) [2022-09-30]. <https://github.com/goto456/stopwords>.

[13] 徐恒,张梦璐,孙德厂. 基于 LDA 模型的国内评论挖掘与情感分析领域主题分析与演化趋势[J]. 河南工业大学学报(社会科学版), 2021, 37(2): 38-46.

[14] Taylor R S. Question-negotiation and information seeking in libraries[J]. C&RL, 1967, 29(3): 88.

[15] 吴晓兵,程晓芸,裴素云,等. 北京市东城区 976 例孕产妇健康信息需求调查[J]. 护理研究, 2020, 34(20): 3715-3718.

[16] 王晓燕,郭睿泽. 孕产妇母婴保健信息需求研究[J]. 科技资讯, 2020, 18(20): 254-256.

[17] 陈浩,郭清,叶成荫,等. 社区妇幼移动健康信息需求分析与思考[J]. 中国初级卫生保健, 2017, 31(11): 38-40, 43.

[18] 时增玉,籍春燕,陶成珍. 妊娠期糖尿病孕产妇信息行为动机体验的质性研究[J]. 中国实用护理杂志, 2021, 37(5): 348-352.

[19] 王晋,支茵. 性别差异视角下网络问答社区减肥信息需求主题特征研究——以“知乎”为例[J]. 现代情报, 2021, 41(2): 89-96, 131.