

基于自然语言处理和深度学习模型的中药黄芪主题建模分析

李建阳¹, 杨移佳¹, 王 骛¹, 彭焕蕊², 曾晓芳¹, 梁京京⁴, 张 伟³, 郑文江^{3*}, 里自然^{2*}

1. 广州中医药大学, 广东 广州 510405

2. 广州医科大学附属中医医院, 广东 广州 510130

3. 广州中医药大学第一附属医院, 广东 广州 510405

4. 湄潭县中西医结合医院 肺病科, 贵州 遵义 564100

摘要: **目的** 基于自然语言处理和深度学习语义建模方法, 分析黄芪文献的研究现状、主题结构及演化特征, 为黄芪研究热点识别和知识结构梳理提供参考。 **方法** 检索 PubMed、Web of Science、Embase、中国知网、万方数据和维普资讯数据库中黄芪相关文献, 检索字段限定为题名和关键词。经 NoteExpress 去重及人工筛选后, 以题名/标题、关键词和摘要构建联合文本, 采用文本清洗、术语规范化和语义嵌入等自然语言处理方法, 结合 PubMedBERT 与 BERTopic 模型进行主题识别、可视化分析和 Topic 0 子主题分析。 **结果** 最终纳入中文文献 12 159 篇, 英文文献 1 536 篇。中文文献 2015 年后发文量明显增加, 英文文献总体呈持续增长趋势。主题建模结果显示, 中文文献形成 19 个主要主题, 主要涉及黄芪免疫调节、肿瘤免疫调控、改善周围神经病变、糖尿病肾病防治等药理作用, 以及黄芪甲苷质量控制和磷脂酰肌醇-3-羟激酶 (phosphatidylinositol-3-hydroxykinase, PI3K)-蛋白激酶 B (protein kinase B, Akt) 通路机制等方向; 英文文献形成 20 个主要主题, 主要涉及黄芪方药机制、质量分析、化学成分以及与癌症、纤维化、缺血再灌注损伤、糖尿病及肠道微生态相关的研究方向。Topic 0 子主题分析显示, 黄芪相关中文文献核心主题呈现“成分-制剂-复方-疾病-机制”交织特征, 英文文献核心主题更突出机制化和模型化表达。 **结论** 黄芪的中英文研究均呈现持续增长和多主题分化趋势, 中文文献更侧重临床应用、复方制剂、糖尿病肾病和质量控制, 英文文献更侧重分子机制、网络药理学、天然产物化学和现代药理验证。自然语言处理与深度学习主题建模方法可有效揭示黄芪研究的知识结构和演化路径, 为后续研究选题和证据整合提供参考。

关键词: 黄芪; 自然语言处理; 深度学习; PubMedBERT; BERTopic; 主题建模; 糖尿病肾病

中图分类号: TP18; G350; R282.71 **文献标志码:** A **文章编号:** 0253-2670(2026)17-6887-14

DOI: 10.7501/j.issn.0253-2670.2026.17.020

Topic modeling analysis of *Astragali Radix* based on natural language processing and deep learning models

LI Jianyang¹, YANG Yijia¹, WANG Ao¹, PENG Huanrui², ZENG Xiaofang¹, LIANG Jingjing⁴, ZHANG Wei³, ZHENG Wenjiang³, LI Ziran²

1. Guangzhou University of Chinese Medicine, Guangzhou 510405, China

2. The Affiliated TCM Hospital of Guangzhou Medical University, Guangzhou 510130, China

3. The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou 510405, China

4. Department of Pulmonology, Meitan Integrated Traditional Chinese and Western Medicine Hospital, Zunyi 564100, China

Abstract: Objective To analyze the research status, thematic structure and evolutionary trends of Chinese and English literature on Huangqi (*Astragali Radix*) using natural language processing and deep-learning-based semantic modeling, and to provide references for hotspot identification and knowledge structure mapping. **Methods** Literature related to *Astragali Radix* was retrieved from PubMed, Web of Science, Embase, China National Knowledge Infrastructure (CNKI), Wanfang Data and VIP databases, with the retrieval fields limited to titles and keywords. After deduplication using NoteExpress and manual screening, titles, keywords and

收稿日期: 2026-05-17

基金项目: 广州市科技局基础研究计划 (2023A03J0306); 广州市科学技术局青年博士“启航”项目 (2025A04J3795); 2026 年广州中医药大学“揭榜挂帅”青年拔尖人才团队项目 (郑文江)

作者简介: 李建阳, 男, 主要研究中西医结合防治呼吸系统疾病的临床和基础研究。E-mail: ljyskywalker@163.com

***通信作者:** 郑文江, 男, 主要从事呼吸系统疾病的临床与基础研究。E-mail: zhengwenjiang92402@gzucm.edu.cn

里自然, 男, 主要从事呼吸系统疾病的中西医结合的临床与基础研究。E-mail: 75306479@qq.com

abstracts were combined to construct the analytical corpus. Natural language processing approaches including text cleaning, terminology normalization and semantic embedding were performed, and PubMedBERT combined with BERTopic was used for topic identification, visualization and subtopic analysis of Topic 0. **Results** A total of 12 159 Chinese publications and 1 536 English publications were included. The publication output of Chinese publications increased markedly after 2015, while English publications showed a continuous upward trend. Topic modeling identified 19 major topics in Chinese literature, mainly involving immunomodulation, diabetic nephropathy, quality control of astragaloside IV, peripheral neuropathy, tumor immunoregulation and phosphatidylinositol-3-hydroxykinase (PI3K)-protein kinase B (Akt)-related network pharmacology. In English literature, 20 major topics were identified, mainly involving formula-related mechanisms, quality analysis of *Astragali Radix*, network pharmacology and cancer, chemical constituents, fibrosis, ischemia-reperfusion injury, diabetes and gut microbiota. Subtopic analysis of Topic 0 suggested that the Chinese core topic showed an intertwined pattern of “components-preparations-formulas-diseases-mechanisms”, whereas the English core topic was more mechanism- and model-oriented. **Conclusion** Both Chinese and English *Astragali Radix* literature showed continuous growth and thematic diversification. Chinese studies focused more on clinical application, compound formulas, diabetic nephropathy and quality control, whereas English studies emphasized molecular mechanisms, network pharmacology, natural product chemistry and modern pharmacological validation. Natural language processing and deep-learning-based topic modeling can effectively reveal the knowledge structure and evolutionary trajectory of *Astragali Radix* research, providing references for subsequent research topic selection and evidence integration.

Key words: *Astragali Radix*; natural language processing; deep learning; PubMedBERT; BERTopic; topic modeling; diabetic nephropathy

黄芪为豆科植物蒙古黄芪 *Astragalus membranaceus* (Fisch.) Bge. var. *mongholicus* (Bge.) Hsiao 或膜荚黄芪 *A. membranaceus* (Fisch.) Bge. 的干燥根,是临床常用补气类中药^[1-3]。现代研究显示,黄芪含有多糖、皂苷、黄酮、氨基酸及多种微量成分,其中黄芪多糖、黄芪甲苷、毛蕊异黄酮、芒柄花素等被认为是其发挥药理作用的重要物质基础^[2-4]。黄芪及其活性成分具有免疫调节、抗炎、抗氧化、抗肿瘤、心血管保护、糖脂代谢调节、肠道屏障保护和器官保护等多重作用^[5-7]。近年来,围绕黄芪多糖、黄芪甲苷及黄芪黄酮类成分的研究不断拓展,相关研究已延伸至糖尿病及其并发症、肿瘤、神经退行性疾病、心血管疾病、肠道疾病和免疫调节等多个方向^[8-10]。随着中药现代化、天然产物药理学、网络药理学、组学技术和人工智能技术的发展,黄芪研究数量持续增加,研究内容由药材鉴定、成分分析和药效评价,逐渐拓展至分子机制、通路调控、复方配伍、临床转化和质量控制等方向^[11-13]。

既往黄芪相关综述多依赖人工阅读和经验归纳,传统文献计量学研究则主要基于关键词共现、作者合作、机构分布和引文关系等信息,虽可揭示研究热点,但对题名、关键词和摘要中的深层语义关系利用不足^[14-15]。黄芪研究中还存在“黄芪”“黄芪甲苷”“黄芪多糖”“*Astragali Radix*”“*Astragalus polysaccharide*”“astragaloside”等多种中英文术语表

和上下文语义关系。自然语言处理可用于文本清洗、术语规范化、命名实体识别、语义表征和文本分类,深度学习预训练语言模型则可进一步捕捉医学文本的上下文语义。基于 transformer 的双向编码器表示 (bidirectional encoder representations from transformers, BERT) 及其生物医学领域衍生模型推动了医学文本挖掘从词袋模型向语义表征模型转变^[16-18]。其中, PubMedBERT 基于生物医学语料进行领域特异性预训练,能够增强医学术语、药物名称、疾病名称和生物学过程的语义表征能力^[16]; BioBERT、SciBERT 和 Sentence-BERT 等模型也为科学文献语义表征、文本相似性计算和下游文献挖掘任务提供了重要方法基础^[18-20]。

在主题建模方面,传统潜在狄利克雷分配 (latent dirichlet allocation, LDA) 等模型长期用于大规模文本潜在主题发现,但其主要依赖词袋表示,对语义上下文和同义表达的捕捉能力有限^[21]。BERTopic 通过预训练语言模型生成文档嵌入,结合统一流形近似与投影 (uniform manifold approximation and projection, UMAP) 降维、基于层次密度的带噪声空间聚类 (hierarchical density-based spatial clustering of applications with noise, HDBSCAN) 及基于类别的词频-逆文档频率 (class-based term frequency-inverse document frequency, c-TF-IDF) 提取主题代表词,可在保留文本语义信息的基础上实现主题识别、主题表示和动态演化

分析^[22-24]。基于此，本研究整合 PubMed、Web of Science、Embase、中国知网、万方数据和维普资讯数据库中 2001—2025 年黄芪相关中英文文献，构建题名/标题、关键词和摘要联合文本，采用自然语言处理与深度学习语义建模方法，结合 PubMedBERT 与 BERTopic 开展主题识别、主题命名、可视化分析和 Topic 0 子主题分析，系统揭示黄芪中英文文献的研究现状、主题结构、热点演化及中外研究差异，以期对黄芪现代研究的知识体系整理、热点识别和后续选题提供参考^[16,22,25]。

1 研究设计

1.1 数据来源与检索策略

本研究数据来源于 PubMed、Web of Science、Embase、中国知网、万方数据和维普资讯数据库。检索时间范围设定为 2001 年 1 月 1 日至 2025 年 12 月 31 日，所有数据库的最终检索于 2026 年 6 月 10 日完成，以尽可能纳入截至 2025 年底已发表并完成数据库收录的文献，2026 年发表的文献

不纳入研究。为提高检索结果的相关性，检索字段限定为题名和关键词，不采用主题扩展检索。英文检索词包括 “*Astragalus membranaceus*” “*Astragalus mongholicus*” “*Astragali Radix*” “*Radix Astragali*” “*Astragalus Radix*” “Huangqi” “Huang Qi” “Huang-qi” “astragaloside” “*Astragalus polysaccharide*” 等；中文检索词包括 “黄芪” “蒙古黄芪” “膜荚黄芪” “黄芪甲苷” “黄芪多糖” 等。英文数据库中，PubMed、Web of Science 和 Embase 的题录数据以 RIS、NBIB、CSV、txt 或数据库支持的 NoteExpress 兼容格式导出；中文数据库中，中国知网、万方数据和维普资讯题录以 NoteExpress 可识别格式、RIS 或 Excel 格式导出。导出字段包括题名/标题、作者、发表年份、来源期刊、摘要、关键词、数据库来源、DOI 或其他可识别信息。所有题录统一导入 NoteExpress 进行格式整理和查重。黄芪中英文文献数据处理与主题建模方法如表 1 所示。

表 1 黄芪中英文文献数据处理与主题建模方法

Table 1 Data processing and topic modeling methods for Chinese and English literature on <i>Astragali Radix</i>		
项目	英文文献	中文文献
数据库来源	PubMed、Web of Science、Embase	中国知网、万方数据、维普资讯
检索范围	2001—2025 年；题名/关键词检索	
数据导出格式	RIS、NBIB、CSV、txt 或 NoteExpress 兼容格式	NoteExpress 格式、RIS 或 Excel
数据整理与筛选	NoteExpress 去重；依据题名、关键词、摘要人工筛选	
分析文本构建	题名+关键词+摘要	
文本预处理	小写化、停用词过滤、词形还原、术语规范化	分词、停用词过滤、同义词规范化、术语词典校正
语义表征与主题建模	PubMedBERT 语义嵌入+BERTopic	BAAI/bge-m3 语义嵌入+BERTopic
主题结构分析	UMAP、主题网络、层次聚类、相似性矩阵、战略坐标图、时间演化	UMAP、层次聚类、相似性矩阵、时间演化
核心主题分析	Topic 0 子主题分析	

1.2 文献纳入与排除标准

1.2.1 纳入标准 ①文献题名、关键词或摘要与黄芪药材、黄芪提取物、黄芪活性成分、黄芪复方应用、黄芪质量控制、黄芪药理机制或临床研究直接相关；②题录信息较完整，至少包含题名/标题、发表年份和来源期刊；③题名、关键词或摘要具有可用于语义建模的有效文本信息；④文献发表时间在 2001—2025 年。

1.2.2 排除标准 ①重复文献；②黄芪仅作为背景性描述或非核心成分出现，无法独立提取主题信息的文献；③主要研究内容为农业种植、畜牧饲养、

植物分类、重金属检测、化工材料或其他与黄芪药理、药学及临床研究无直接关系的文献；④会议纪要、新闻报道、书评、专利、编辑材料、学位论文及非正式学术记录；⑤摘要为空、题录严重缺失或文本过短、无法进行主题建模的记录。

1.3 数据去重、人工筛选与联合文本构建

各数据库题录首先统一导入 NoteExpress 进行格式整理和去重。去重时综合参考题名、作者、发表年份、期刊来源、DOI 和数据库编号等信息；对于题名高度相似但来源期刊名称存在大小写、缩写或中英文表达差异的记录，结合作者、年份和摘要

内容进行人工复核。去重完成后,进一步依据题名、关键词和摘要进行人工筛选,排除非黄芪核心主题、非药理药学相关、题录信息不完整或摘要语义信息不足的记录。为提高主题建模的语义信息密度,本研究将每篇文献的题名/标题、关键词和摘要合并为联合文本。英文文献采用“title+keywords+abstract”构建分析文本,中文文献采用“标题+关键词+摘要”构建分析文本。该处理方式既保留题名和关键词中的高信息量术语,又利用摘要中的上下文描述增强语义建模效果。

1.4 文本预处理与自然语言处理流程

基于 Python 完成文本预处理。首先对题名/标题、关键词和摘要字段进行完整性检查,剔除空值、文本过短、仅含标点符号、无实际语义内容及标注为“nan”“null”等无效文本的记录,并将题名/标题、关键词和摘要合并为联合文本。英文文本主要进行小写化、停用词过滤、词形还原和术语规范化处理;中文文本主要进行分词、停用词过滤、同义词规范化和专业术语词典校正。上述处理流程参考生物医学自然语言处理中文本规范化、语义表示和领域文本挖掘的一般方法,以减少无效词项、同义表达和非专业噪声对后续主题识别的影响^[26-27]。为减少专业术语误拆分,本研究结合黄芪药理研究特点构建自定义术语词典,对“黄芪甲苷”“黄芪多糖”“糖尿病肾病”“磷脂酰肌醇-3-羟激酶(phosphatidylinositol-3-hydroxykinase, PI3K)/蛋白激酶 B(protein kinase B, Akt)”“网络药理学”等词进行规范化处理。特征工程阶段采用词项向量化方法构建文本特征,并结合 c-TF-IDF 提取主题代表词,设置 ngram_range=(1, 1)、最大特征数为 8 000、最小文档频率为 3,以降低低频噪声词对主题识别的影响^[22,26]。

1.5 PubMedBERT 语义嵌入与 BERTopic 主题建模

深度学习模型主要用于文献联合文本的语义表征。本研究在完成文本清洗和术语规范化后,采用预训练语言模型对题名/标题、关键词和摘要构建的联合文本进行语义编码,生成文献级语义嵌入向量。英文文献主要采用 PubMedBERT 进行生物医学语义表征,以增强医学术语、药理机制、疾病名称和生物学过程的上下文表达能力^[16];中文文献则在中文分词和专业术语规范化基础上,采用 BAAI/bge-m3 多语言预训练语义嵌入模型进行文本向量化。该模型适用于中文及多语言语义表示,可将中文联合文本编码为文献级语义嵌入向量。语义编码

时批处理大小设置为 32,最大序列长度设置为 512,对超过最大长度的文本进行截断处理,并对输出向量进行归一化。随后采用 BERTopic 进行主题建模,即通过 UMAP 对高维语义向量进行降维,利用 HDBSCAN 对语义相近文献进行聚类,并采用 c-TF-IDF 提取各主题代表词^[22-24]。模型设置为自动优化主题数量, nr_topics=“auto”,最小主题样本量设置为 min_topic_size=5,并开启主题概率计算;建模完成后导出主题编号、主题文献量、主题关键词、文献主题分配和主题概率等结果。所有建模过程均设置固定随机种子 42,以提高结果可重复性。

1.6 主题可视化、时间演化与 Topic 0 子主题分析

为系统展示黄芪中英文文献的主题结构和演化特征,本研究构建多维度可视化体系。首先绘制年度发文趋势图和来源期刊排名图,以反映文献增长和来源分布;随后绘制 UMAP 语义景观图和主题关键词贡献图,用于展示文献在二维语义空间中的聚类分布及各主题的代表性词项^[22-23]。主题间结构关系通过层次聚类图、主题相似性矩阵、主题语义网络图和战略坐标图进行展示,其中主题相似性采用余弦相似度计算,层次聚类采用 Ward 链接法构建。为分析主题动态变化,结合文献发表年份统计各主题年度文献量、主题占比或标准化活跃度,绘制主题时间演化图、年度趋势图和主题活跃度热图。针对中英文文献中的核心主题 Topic 0,进一步提取其内部文献语义嵌入向量进行子主题分析,并绘制子主题语义分布图、关键词贡献图、层次聚类图和年度趋势图,以揭示核心主题内部的结构分化和时间演化特征^[22,28]。

2 结果与分析

2.1 文献筛选流程与数据集构建

本研究围绕 2001—2025 年黄芪相关中英文文献开展多数据库检索。经 NoteExpress 去重后,英文文献保留 4 321 篇,中文文献保留 14 134 篇;进一步依据题名、关键词和摘要进行人工筛选后,最终英文文献纳入 1 536 篇,中文文献纳入 12 159 篇。需要说明的是,最终纳入文献为题名/关键词限定、多数据库去重及人工相关性筛选后形成的高相关分析集。筛选过程中排除了农业种植、畜牧饲养、植物分类、重金属检测、化工材料及题录信息不足等与黄芪药理药学和临床研究相关性较低的记录,因此,最终纳入量低于宽口径初检结果,但提高了主题识别的相关性和语义聚类稳定性。随后将题名/标

题、关键词和摘要构建为联合文本，用于后续 PubMedBERT 语义嵌入、BERTopic 主题建模及可

视化分析。文献检索、去重、筛选、联合文本构建和主题建模流程见图 1。



图 1 黄芪中英文文献检索、筛选与主题建模流程

Fig. 1 Workflow of retrieval, screening and topic modeling of Chinese and English literature on *Astragali Radix*

2.2 黄芪中英文文献的年度发文趋势与来源分布

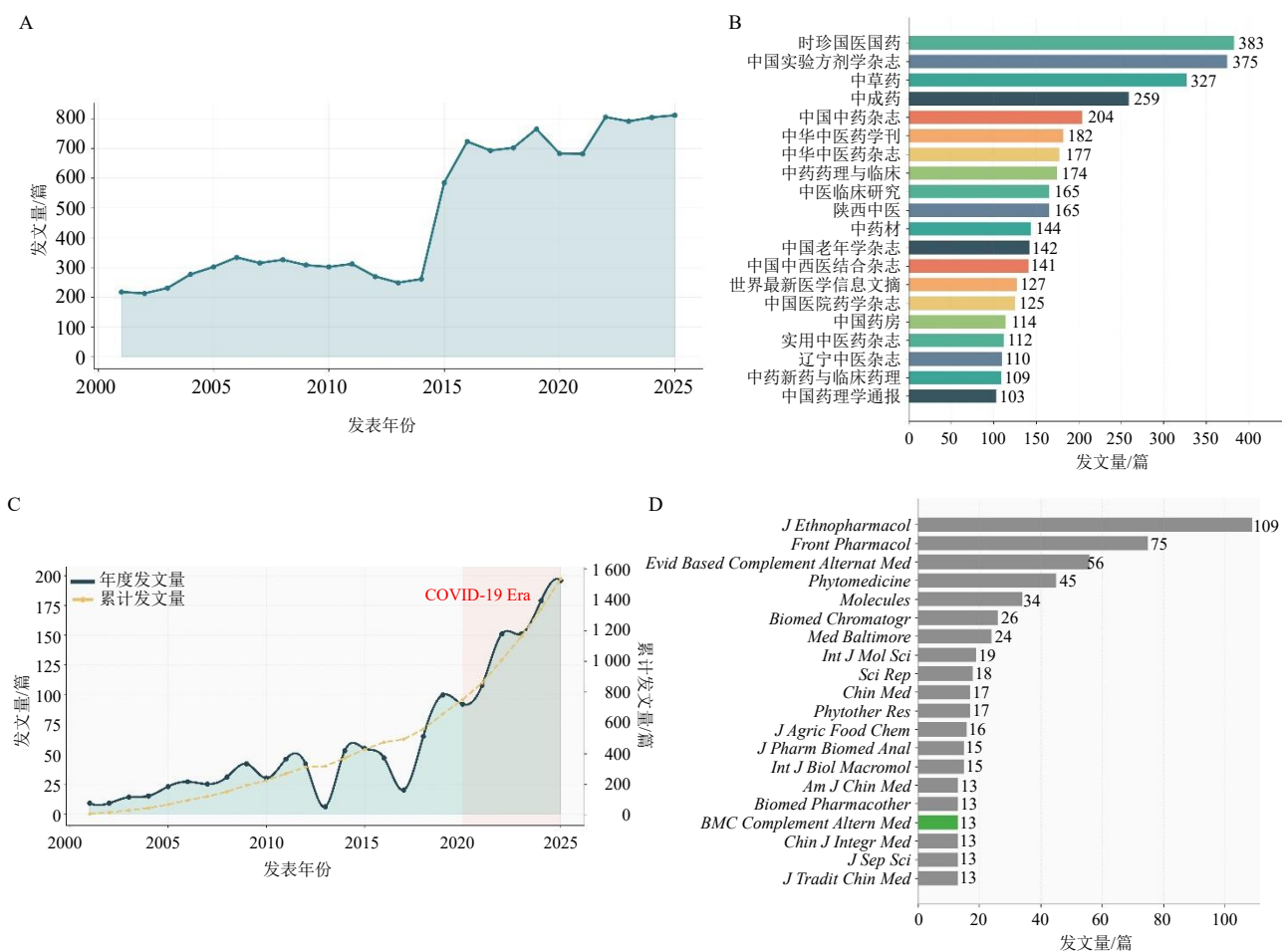
黄芪相关中文文献年度发文趋势（图 2-A）显示，2001—2014 年发文量总体处于缓慢增长阶段，年发文量在 200~300 篇波动；2015 年后发文量明显跃升，提示黄芪研究在中文数据库中进入快速增长阶段。2022—2025 年发文量维持在高位，说明黄芪仍是中医药研究中的持续热点。来源期刊分布显示，黄芪中文文献主要集中于中医药和中药学相关期刊（图 2-B），其中《时珍国医国药》《中国实验方剂学杂志》《中草药》《中成药》《中国中药杂志》等发文量居前，反映黄芪研究在中药药理、方剂学、质量控制和临床应用领域具有较高关注度。

英文文献年度发文量与累积发文量整体呈上升趋势（图 2-C），尤其近年增长较明显，提示黄芪研究的国际关注度逐渐提高。英文来源期刊主要集中于民族药理学、现代药理学、天然产物化学和补充替代医学领域（图 2-D），其中 *Journal of Ethnopharmacology*、*Frontiers in Pharmacology*、

Evidence-Based Complementary and Alternative Medicine、*Phytomedicine* 和 *Molecules* 等发文量居前。中英文来源期刊差异说明，中文研究更多依托中医药专业期刊体系，英文研究则更多进入国际药理学、天然产物和综合医学期刊平台。

2.3 黄芪中英文文献主题识别与语义空间分布

基于 BERTopic 的主题识别结果显示，黄芪中文文献共形成 19 个主要主题。UMAP 语义景观图（图 3-A）显示，不同主题在二维语义空间中形成多个相对集中的语义簇，提示模型能够根据文献内容识别出较稳定的研究方向。中文主题关键词贡献图（图 3-B）显示，主要主题涉及黄芪免疫调节、糖尿病肾病、黄芪甲苷质量控制、糖尿病周围神经病变、胃溃疡与脾胃虚寒、黄芪多糖与肿瘤免疫以及 PI3K-Akt 与糖尿病肾病等方向。糖尿病肾病-PI3K/Akt-网络药理学等方向。整体来看，中文文献主题既包括黄芪基础药理机制，也包括复方应用、注射液、糖尿病肾病、质量控制和临床相关研究。



A-中文文献发文趋势；B-中文文献来源期刊；C-英文文献发文趋势；D-英文文献来源期刊。

A-publication trend of Chinese literature; B-source journals of Chinese literature; C-publication trend of English literature; D-source journals of English literature.

图 2 黄芪中英文文献发文趋势与来源分布

Fig. 2 Publication trends and source distributions of Chinese and English literature on *Astragali Radix*

黄芪英文文献共形成 20 个主要主题。UMAP 语义景观图 (图 3-C) 显示, 英文文献主题在语义空间中呈现多个相对独立但又局部关联的聚类区域。主题关键词贡献图 (图 3-D) 显示, 英文文献主要围绕黄芪方药与通路机制、质量分析, 网络药理学与癌症、肿瘤细胞、化学成分与发酵、纤维化、缺血再灌注损伤、生物合成、当归补血汤、糖尿病及肠道微生态等方向展开。与中文文献相比, 英文研究更突出机制解释、药理模型、网络药理学、天然产物化学和疾病机制验证 (表 2)。

BERTopic 聚类过程中, 部分文献因文本信息密度较低、主题边界不清或同时涉及多个研究方向被归入噪声主题。经统计, 中文文献噪声文献 2 606 篇, 约占中文建模文献的 21.8%; 英文文献噪声文

献 319 篇, 约占英文建模文献的 20.8%。结合题名和摘要内容初步核查发现, 噪声文献主要表现为研究对象表述不够集中、摘要信息量不足, 或同时涉及药材、复方、疾病模型、活性成分和机制指标等多个方向。上述文献未参与主题命名和主题结构归纳, 但保留于 UMAP 语义分布图中用于展示整体语义空间。由于中英文主要主题仍形成相对稳定的语义簇, 且有效主题能够覆盖药效机制、质量控制、临床应用、网络药理学和疾病模型等核心方向, 提示噪声文献未对全局主题结构判断产生明显影响。

2.4 黄芪中文文献研究主题的层次结构与相似性

黄芪中文文献主题层次聚类图 (图 4) 显示, 19 个主题之间并非完全孤立, 而是在语义距离上形成多个相对稳定的主题群。其中, 黄芪免疫调节、



A-中文文献语义分布; B-中文文献主题词贡献; C-英文文献语义分布; D-英文文献主题词贡献; 灰色点为噪声文献, 中文 2 606 篇, 占 21.8%, 英文 319 篇, 占 20.8%。

A-semantic distribution of Chinese literature; B-contribution of subject terms in Chinese literature; C-semantic distribution of English literature; D-contribution of subject terms in English literature; gray dots represent noisy literature, including 2 606 Chinese-language papers accounting for 21.8%, and 319 English-language papers accounting for 20.8%.

图 3 黄芪中英文文献语义分布与主题词贡献

Fig. 3 Semantic distribution and topic word contribution of Chinese and English literature on *Astragali Radix*

糖尿病肾病、PI3K-Akt 相关研究、细胞凋亡等主题之间具有较近语义距离, 提示中文黄芪研究中“黄

芪-免疫功能-炎症/凋亡-糖尿病肾病”构成重要研究链条。质量控制、黄芪甲苷含量测定和 HPLC 分析

表 2 黄芪中英文文献主题结构差异概括

维度	中文文献	英文文献
主题数量	形成 19 个主要主题	形成 20 个主要主题
核心研究方向	免疫调节、糖尿病肾病、黄芪甲苷质量控制、方药与通路机制、质量分析、网络药理学与癌症、周围神经病变、胃溃疡、肿瘤免疫调控、PI3K-Akt 网络药理学等	肿瘤细胞、化学成分与发酵、纤维化、缺血再灌注损伤、生物合成、糖尿病及肠道微生态等
研究切入点	更强调临床应用、复方配伍、注射液、中成药、疾病治疗和质量控制	更强调分子机制、网络药理学、天然产物化学、疾病模型和现代药理验证
代表性疾病或应用场景	糖尿病肾病、周围神经病变、胃溃疡、肿瘤免疫、心肌炎等	癌症、糖尿病、纤维化、缺血再灌注损伤、肠道微生态等
整体特征	体现“临床应用-复方制剂-质量评价-机制探索”并行发展的特点	体现“成分分析-分子机制-疾病模型-国际化药理验证”逐步深化的特点

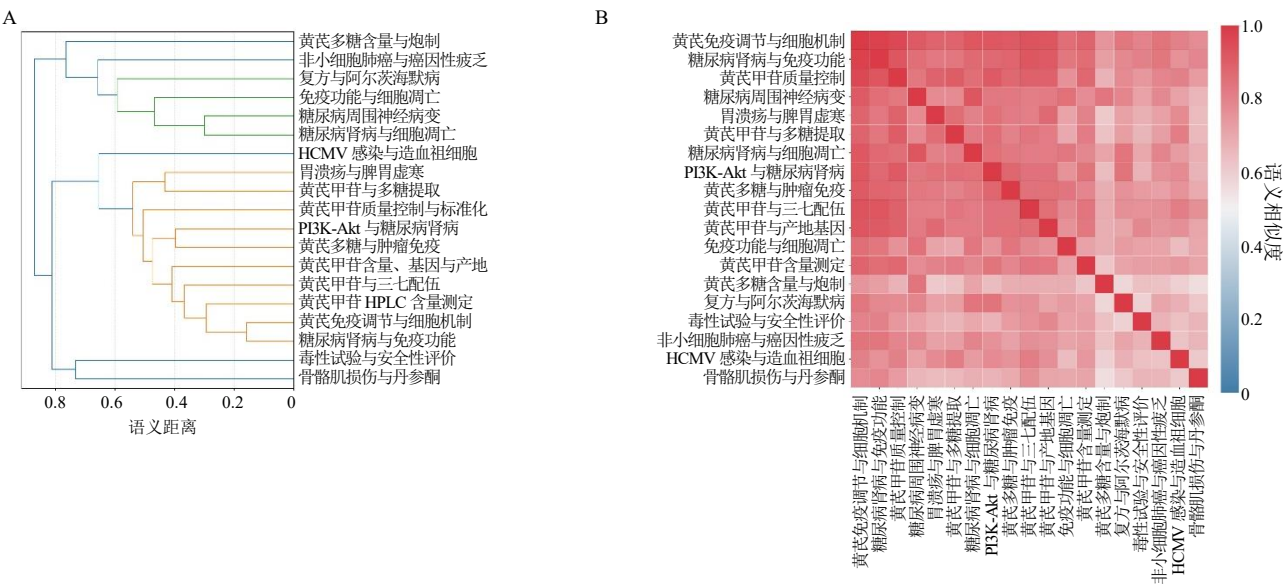


图 4 黄芪中文文献主题层次聚类 (A) 与语义相似性矩阵 (B)

Fig. 4 Hierarchical clustering (A) and semantic similarity matrix (B) of topics in Chinese literature on *Astragali Radix*

等主题则形成相对独立的药学基础研究分支，反映黄芪中文研究不仅关注药效机制，也重视中药质量评价和成分控制。

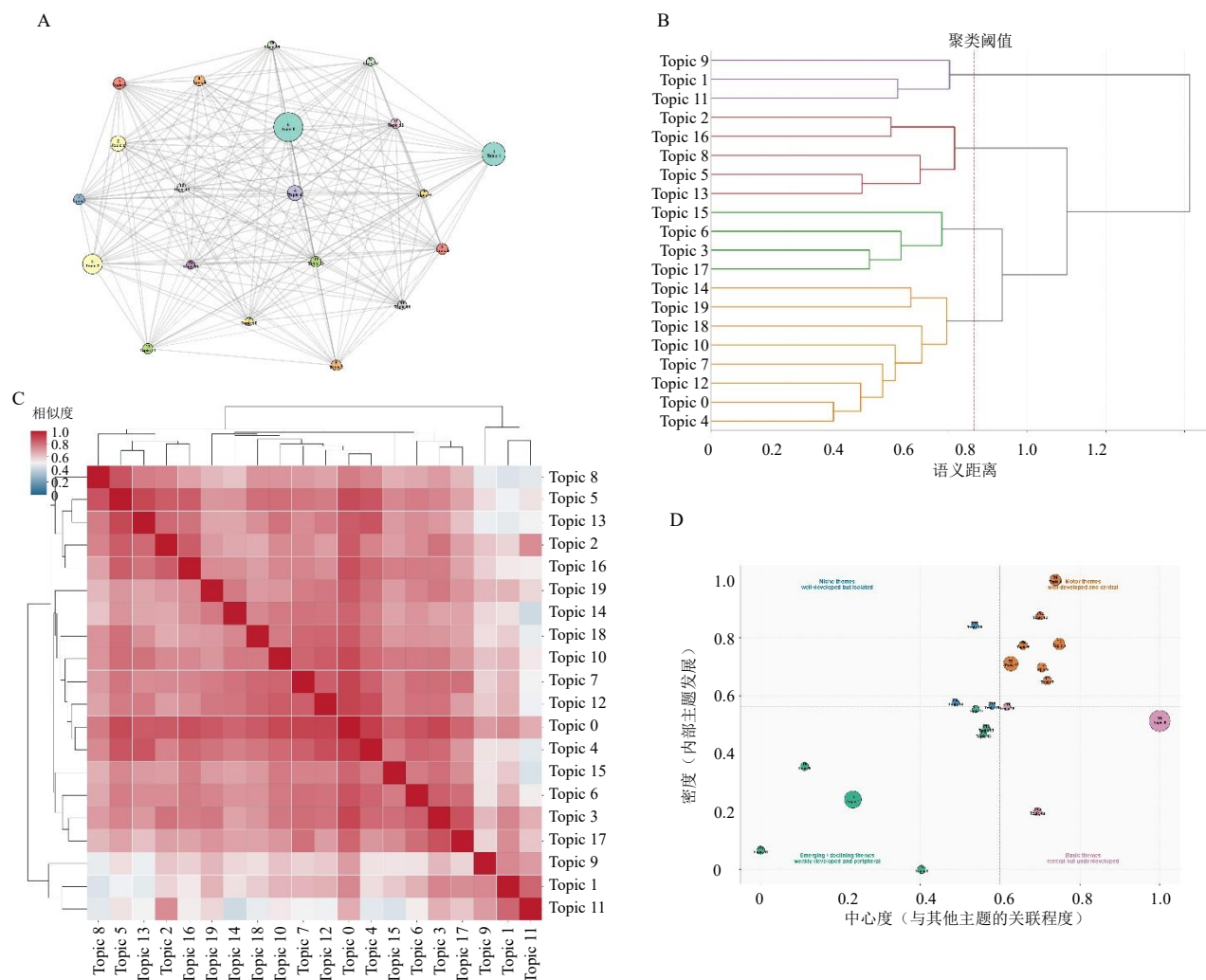
主题语义相似性矩阵进一步显示，多数中文主题之间存在中高强度语义关联，尤其是免疫功能、糖尿病肾病、PI3K-Akt 相关研究、肿瘤免疫和黄芪甲苷相关主题之间相似度较高。这说明黄芪中文研究常以疾病应用为牵引，将药理机制、免疫调节、网络药理学和质量控制结合起来。相比之下，粒细胞-巨噬细胞集落形成单位 (granulocyte-macrophage colony-forming unit, CFU-GM) 与人巨细胞病毒 (human cytomegalovirus, HCMV) 相关研究、骨髓肌、运动干预等主题与其他主题的相似度相对较

低，提示其可能属于较为独立或阶段性发展的研究方向。

2.5 黄芪英文文献研究主题的语义网络、层次结构与战略分布

英文黄芪文献主题语义网络图 (图 5-A) 显示，不同主题之间存在复杂关联，其中 Topic 0 和 Topic 1 节点相对较大，提示方药与通路机制、质量分析等方向在英文文献中具有较高主题规模和连接度。主题网络中的多条连线表明，英文黄芪研究并非单一方向发展，而是在药理机制、质量评价、网络药理学、癌症、糖尿病、肠道微生态和天然产物成分之间形成交叉联系。

层次聚类结果 (图 5-B) 显示，英文主题可大



A-语义网络; B-层次聚类; C-相似性热图; D-战略坐标图。

A-semantic network; B-hierarchical clustering; C-similarity heatmap; D-strategic diagram.

图 5 黄芪英文文献主题结构分析

Fig. 5 Topic structure analysis of English literature on *Astragali Radix*

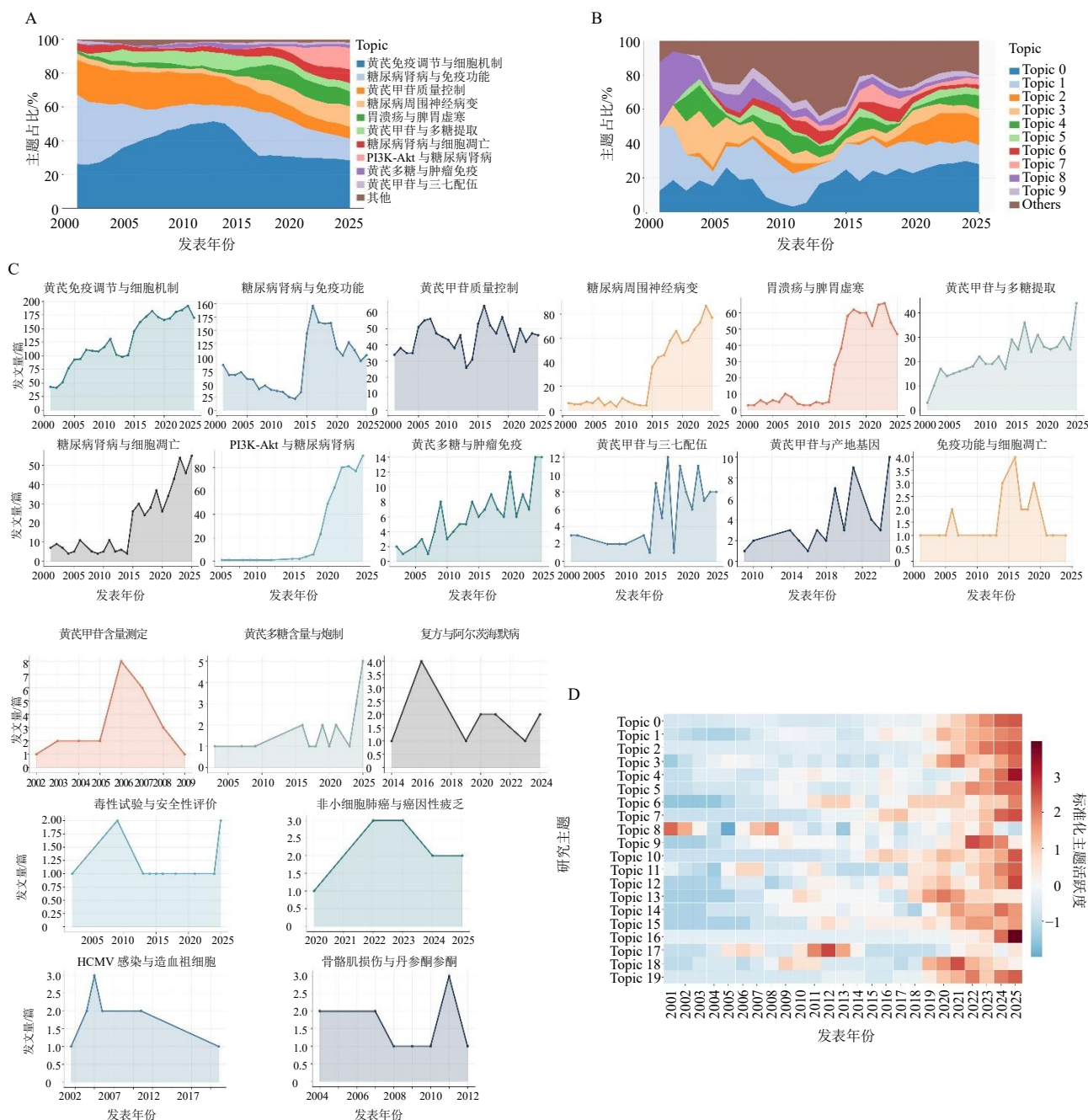
致分为几个语义群：一类围绕质量分析、黄芪、质谱和成分鉴定展开；一类围绕网络药理学、癌症、肿瘤细胞和分子靶点展开；一类围绕糖尿病、肠道微生态、缺血再灌注损伤和器官保护展开；另有部分主题涉及当归补血汤、生物合成、发酵和抗氧化等方向。相似性热图（图 5-C）进一步提示，英文文献中多数主题之间存在一定交叉，但不同主题群仍具有较好区分度。

战略坐标图（图 5-D）从中心性和密度 2 个维度揭示了英文主题的发展状态。部分主题位于“motor themes”区域，说明其既具有较强内部发展程度，又与其他主题保持较高连接度，是当前黄芪英文研究中发展较成熟且具有带动作用的方向。

Topic 0 位于中心性较高但密度中等的位置，提示其具有广泛联系，但内部仍有继续深化空间；部分主题位于“emerging/declining themes”区域，提示其可能处于新兴发展或热度下降阶段，需要结合时间演化结果进一步判断。

2.6 黄芪中英文研究主题的时间演化特征

黄芪中文研究主题时间演化图（图 6-A）显示，黄芪免疫调节基础机制长期占据较大比例，是中文黄芪文献中最稳定的核心方向。糖尿病肾病、黄芪甲苷质量控制、胃溃疡、PI3K-Akt 相关研究和肿瘤免疫等主题随时间逐渐扩展，提示黄芪中文研究从早期经验应用和基础药效观察，逐步转向疾病机制、靶点预测、质量控制和临床相关研究。各主题年度趋势



A-中文研究主题演化; B-英文研究主题演化; C-中文研究年度趋势; D-英文研究主题活跃度。

A-evolution of Chinese research topics; B-evolution of English research topics; C-annual trends of Chinese research; D-activity level of English research topics.

图 6 黄芪中英文文献主题演化趋势

Fig. 6 Topic evolution trends of Chinese and English literature on *Astragali Radix*

图(图 6-C)进一步显示, 2015 年后多个主题同步增长, 与黄芪中文文献整体发文章量跃升相一致。

黄芪英文研究主题时间演化图(图 6-B)显示, 英文文献中 Topic 0、Topic 1、Topic 2 等主要主题在不同时期占据不同权重。早期英文研究更多集中于方药、天然产物成分和质量分析, 随

后逐步扩展至网络药理学、癌症、糖尿病、肠道微生物、纤维化和缺血再灌注损伤等方向。标准化主题活跃度热图(图 6-D)显示, 多个英文主题在 2020 年后活跃度明显升高, 提示国际黄芪研究正在向多疾病场景、多靶点机制和现代药理学验证方向拓展。

综合来看,中英文黄芪研究的时间演化均表现出“核心主题持续存在,新兴主题不断出现”的特征。中文研究的增长更集中于临床应用、复方、注射液、糖尿病肾病和质量控制;英文研究则更强调分子机制、网络药理学、肿瘤、天然产物化学和疾病模型验证。这种差异反映了中英文文献在研究传统、发表平台和问题意识上的不同。

3 核心主题 Topic 0 的子主题分析

3.1 黄芪中文文献 Topic 0 的子主题结构

中文文献 Topic 0 子主题语义分布图(图 7-A)显示,核心主题内部进一步分化出多个子主题区域,提示“黄芪-免疫调节-基础机制”并非单一研究方向,而是由多类研究内容共同构成的复合型主题。子主题关键词贡献结果(图 7-B)显示,Topic 0 内部涉及注射液、黄芪甲苷、黄芪多糖、大鼠模型、表达调控、血管生成、心肌炎、建中汤、黄芪总黄酮(total flavonoids of *Astragalus*, TFA)、抗病毒、辐射损伤、脑源性神经营养因子(brain derived neurotrophic factor, BDNF)和运动干预等方向。该结果说明,黄芪中文核心主题具有明显的“成分-制剂-复方-疾病-机制”交织特征,既包括黄芪活性成分和制剂研究,也涵盖动物实验、细胞实验、疾病模型和临床相关应用。

进一步结合子主题层次聚类结果(图 7-C)可见,部分子主题之间具有较近语义距离。例如,黄芪甲苷、黄芪多糖、动物实验模型和注射液等方向联系较为紧密,说明中文黄芪研究常以具体活性成分或制剂为切入点,通过动物或细胞模型观察免疫调节、炎症反应、氧化损伤和组织保护等药理作用。与此同时,心肌炎、病毒抑制、血管生成、运动干预等子主题相对分散,提示 Topic 0 内部还存在若干疾病场景化或应用拓展型研究方向。中文 Topic 0 的复杂结构反映出黄芪在中文研究体系中不仅是一味单一药材,更是贯穿复方应用、成分研究、制剂开发和疾病干预的综合研究对象。

3.2 黄芪英文文献 Topic 0 的子主题组成与年度趋势

英文文献 Topic 0 子主题组成图(图 8-A)显示,其内部 5 个子主题在不同年份所占比例存在动态变化。早期由于黄芪英文文献数量相对有限,Topic 0 内部子主题组成波动较大;随着研究数量各子主题逐渐形成较为稳定的结构。3 年移动平均趋势显示,近年多个子主题均呈上升趋势,说明英文黄芪核心主题内部正在持续扩展,相关研究由早

期方药和药效观察逐渐转向机制解释、活性成分研究、疾病模型验证和多靶点调控。

从研究内容看,英文 Topic 0 更强调机制化和模型化表达。其子主题虽然数量少于中文 Topic 0,但结构相对集中,主要围绕黄芪相关机制研究的组成变化和年度增长展开。与中文文献中大量出现的注射液、复方、临床疾病场景和中成药应用不同,英文文献更倾向于使用标准化的现代药理学表达方式,对黄芪活性成分、分子靶点、通路调控和疾病模型进行归纳。这说明黄芪英文研究更突出“成分-靶点-机制-疾病模型”的研究路径。

4 讨论

本研究的主要方法学价值在于,将自然语言处理、领域预训练语言模型和语义主题建模引入黄芪中英文文献知识结构分析。既往文献计量研究多依赖关键词、作者、机构和引文等外部题录信息,能够反映研究产出和合作关系,但对题名、关键词和摘要中隐含的语义结构挖掘不足^[14-15]。黄芪相关研究涉及药材名称、活性成分、复方制剂、疾病名称、信号通路和实验模型等多层术语,且中英文表达差异明显,仅依赖词频统计容易造成语义割裂。PubMedBERT 基于生物医学语料进行领域特异性预训练,可更好识别医学术语、药物名称和生物学过程之间的上下文关系^[16-18];BERTopic 则在语义嵌入基础上结合 UMAP、HDBSCAN 和 c-TF-IDF,实现文献聚类与主题表示的统一^[22-24]。因此,本研究并非单纯统计黄芪文献数量,而是从语义层面刻画黄芪研究的主题结构、主题间关系及动态演化,为中药复杂文献体系的自动化知识发现提供了可参考的技术路径^[25,29]。

从主题结构看,黄芪中英文研究均围绕活性成分、质量评价、药理机制和疾病干预形成了相对稳定的知识框架,但二者的研究取向存在差异。中文文献更突出黄芪在中医临床实践、复方配伍、制剂应用、糖尿病肾病和质量控制中的研究价值,体现出“药材-复方-疾病-应用”的问题导向;英文文献则更强调黄芪的化学成分、网络药理学、分子靶点、肿瘤、糖尿病、纤维化、缺血再灌注损伤和肠道微生态等方向,体现出“成分-靶点-机制-模型”的研究逻辑^[2-4]。这一差异与中英文研究传统、期刊体系和学科话语方式密切相关:中文研究更重视黄芪在中医药临床体系中的实际应用和方药经验转化,英文研究则更倾向于将黄芪纳入天然产物药理学和



A-语义分布；B-关键词贡献；C-层次聚类；D-年度趋势。

A-semantic distribution; B-keyword contribution; C-hierarchical clustering; D-annual trend.

图7 黄芪中文文献 Topic 0 子主题结构与演化

Fig. 7 Subtopic structure and evolution of Topic 0 in Chinese literature on *Astragali Radix*

代生物学机制验证框架。Topic 0 子主题分析进一步说明，黄芪中文研究核心主题内部呈现成分、制

剂、复方、疾病和机制交织的复合结构，而英文核心主题相对集中于机制化和模型化表达。由此提

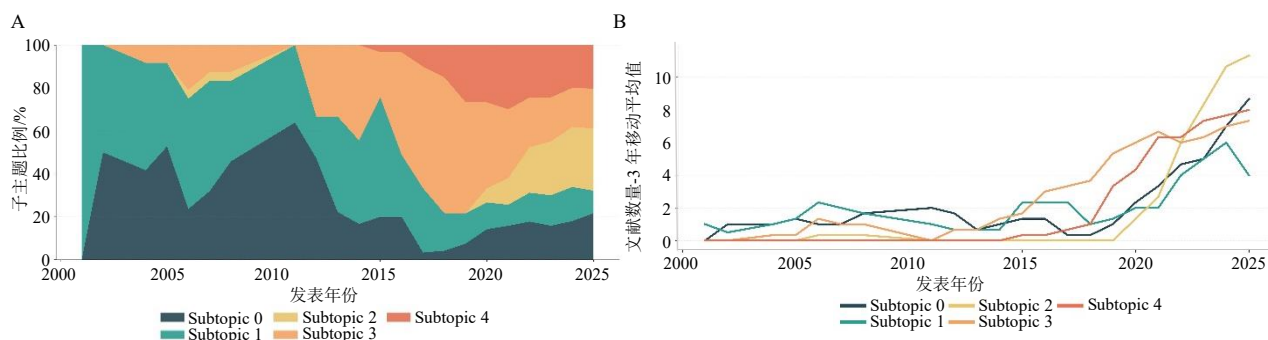


图 8 英文黄芪文献 Topic 0 子主题组成 (A) 与年度变化趋势 (B)

Fig. 8 Subtopic composition (A) and annual trends (B) of Topic 0 in English literature on *Astragali Radix*

示, 黄芪研究未来需要在保留中医药应用特色的基础上, 加强标准化术语表达、机制验证和国际化学术语转化, 使临床经验、物质基础和分子机制之间形成更紧密的证据链^[11,13,30]。

从研究展望看, 黄芪研究应进一步从“主题扩张”转向“证据整合与机制深化”。一方面, 应围绕黄芪多糖、黄芪甲苷、黄芪黄酮等核心物质基础, 加强质量标志物、药效物质群、剂量-效应关系和体内过程研究, 避免仅停留在成分罗列或单一通路预测层面^[6-8]。另一方面, 应结合糖尿病肾病、肿瘤、心血管疾病、肠道疾病、神经退行性疾病和组织修复等优势疾病场景, 推动网络药理学预测、组学分析、实验验证和临床证据之间的闭环整合^[9,11-12]。本研究仍存在一定局限: 其一, 不同数据库的题录格式、摘要完整性和关键词标注方式不同, 可能影响主题分布; 其二, 研究主要基于题名/标题、关键词和摘要构建联合文本, 尚未纳入全文语义信息; 其三, BERTopic 属于无监督主题模型, 主题数量、主题归并和主题命名仍需结合专业判断。后续研究可进一步引入全文数据、引文网络、知识图谱、多模态图表信息和专家共识评估, 提升主题识别的稳定性、可解释性和中医药知识转化价值^[25,28-29]。

5 结语

本研究基于 2001—2025 年黄芪中英文文献, 采用 PubMedBERT 与 BERTopic 相结合的语义主题建模方法, 揭示了黄芪研究的主题结构、时间演化及中英文研究差异。结果表明, 黄芪研究已形成多主题并行发展的知识格局, 中文文献更偏向临床应用、复方制剂、糖尿病肾病和质量控制, 英文文献更偏向分子机制、网络药理学、天然产物化学和现代药理验证。该方法能够从大规模文献中识别潜在

研究主题和演化路径, 为黄芪研究热点判断、知识体系梳理和后续选题提供参考。未来应进一步推动黄芪研究在物质基础、机制验证、质量评价和临床证据之间的融合, 以促进黄芪现代研究的系统化、标准化和国际化发展。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] Liu Y X, Song X M, Dan L W, *et al.* *Astragali Radix*: Comprehensive review of its botany, phytochemistry, pharmacology and clinical application [J]. *Arch Pharmacol Res*, 2024, 47(3): 165-218.
- [2] Wang P P, Wang Z, Zhang Z P, *et al.* A review of the botany, phytochemistry, traditional uses, pharmacology, toxicology, and quality control of the *Astragalus membranaceus* [J]. *Front Pharmacol*, 2023, 14: 1242318.
- [3] Zhang Y Y, Chen Z J, Chen L P, *et al.* *Astragali Radix* (Huangqi): A time-honored nourishing herbal medicine [J]. *Chin Med*, 2024, 19(1): 119.
- [4] Jin X J, Zhang H J, Xie X R, *et al.* From traditional efficacy to drug design: A review of *Astragali Radix* [J]. *Pharmaceuticals*, 2025, 18(3): 413.
- [5] Li S, Hu X P, Liu F, *et al.* Bioactive components and clinical potential of *Astragalus* species [J]. *Front Pharmacol*, 2025, 16: 1585697.
- [6] Liang Y T, Chen B Q, Liang D, *et al.* Pharmacological effects of astragaloside IV: A review [J]. *Molecules*, 2023, 28(16): 6118.
- [7] Wang B, Wu B L, Ma Y J, *et al.* *Astragalus* polysaccharides: Structure-immunomodulation relationships, multi-target pharmacological activities, and cutting-edge applications in immune modulation [J]. *Front Immunol*, 2025, 16: 1714898.
- [8] Shi Y S, Ma P. Pharmacological effects of *Astragalus* polysaccharides in treating neurodegenerative diseases [J]. *Front Pharmacol*, 2024, 15: 1449101.
- [9] Liu S Y, Wang L Y, Zhang Z H, *et al.* The potential of

- Astragalus* polysaccharide for treating diabetes and its action mechanism [J]. *Front Pharmacol*, 2024, 15: 1339406.
- [10] Yang Q, Meng D D, Zhang Q Y, *et al.* Advances in research on the anti-tumor mechanism of *Astragalus* polysaccharides [J]. *Front Oncol*, 2024, 14: 1334915.
- [11] Wei X N, Leng X H, Liang J W, *et al.* Pharmacological potential of natural medicine *Astragali Radix* in treating intestinal diseases [J]. *Biomed Pharmacother*, 2024, 180: 117580.
- [12] Liu M F, Di Y M, May B, *et al.* Renal protective effects and mechanisms of *Astragalus membranaceus* for diabetic kidney disease in animal models: An updated systematic review and Meta-analysis [J]. *Phytomedicine*, 2024, 129: 155646.
- [13] Liu P Y, Hong K F, Liu Y D, *et al.* Total flavonoids of *Astragalus* protects glomerular filtration barrier in diabetic kidney disease [J]. *Chin Med*, 2024, 19(1): 27.
- [14] Chen Y J, Xie M R, Zhou S Q, *et al.* Research state of the herbal medicine Huangqi (*Radix Astragali*): A global and bibliometric study [J]. *Medicine*, 2024, 103(8): e37277.
- [15] Zheng Y J, Ren W Y, Zhang L N, *et al.* A review of the pharmacological action of *Astragalus* polysaccharide [J]. *Front Pharmacol*, 2020, 11: 349.
- [16] Gu Y, Tinn R, Cheng H, *et al.* Domain-specific language model pretraining for biomedical natural language processing [J]. *ACM Trans Comput Healthcare*, 2022, 3(1): 1-23.
- [17] Devlin J, Chang M W, Lee K, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding [A] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [C]. Minneapolis: Minnesota Association for Computational Linguistics, 2019: 4171-4186.
- [18] Lee J, Yoon W, Kim S, *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining [J]. *Bioinformatics*, 2020, 36(4): 1234-1240.
- [19] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text [A] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) [C]. Hong Kong: ACL, 2019: 3613-3618.
- [20] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks [A] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) [C]. Hong Kong: Association for Computational Linguistics, 2019: 3982-3992.
- [21] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. *J Machine Learning Res*, 2003, 3: 993-1022.
- [22] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [J/OL]. *arXiv* (2022-03-11) [2026-07-07]. <https://arxiv.org/abs/2203.05794>.
- [23] McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction [J/OL]. *arXiv* (2018-02-09) [2026-07-07]. <https://arxiv.org/abs/1802.03426>.
- [24] McInnes L, Healy J, Astels S. Hdbscan: Hierarchical density based clustering [J]. *J Open Source Softw*, 2017, 2(11): 205.
- [25] Lyu D M, Wang X B, Chen Y, *et al.* Language model and its interpretability in biomedicine: A scoping review [J]. *iScience*, 2024, 27(4): 109334.
- [26] Chen Q Y, Hu Y, Peng X Q, *et al.* Benchmarking large language models for biomedical natural language processing applications and recommendations [J]. *Nat Commun*, 2025, 16: 3280.
- [27] Usuyama N, Wong C, Zhang S, *et al.* Biomedical natural language processing in the era of large language models [J]. *Annu Rev Biomed Data Sci*, 2025, 8: 471-490.
- [28] Blei D M, Lafferty J D. Dynamic topic models [A] // Proceedings of the 23rd International Conference on Machine Learning-ICML '06 [C]. Pittsburgh: ACM, 2006: 113-120.
- [29] Ma L, Chen R, Ge W G, *et al.* AI-powered topic modeling: Comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women [J]. *Exp Biol Med*, 2025, 250: 10389.
- [30] Han X, Huang L M, Li G, *et al.* Effect of *Astragalus* injection on left ventricular remodeling in HFmrEF: A systematic review and Meta-analysis [J]. *Front Cardiovasc Med*, 2024, 11: 1374114.

[责任编辑 潘明佳]