

多源数据融合下医护科研人员科研能力提升策略研究^{*}

杨晓雯¹ 季汉珍²

(¹ 南京医科大学图书馆 南京 211166

² 南通市第三人民医院/南通大学附属南通第三医院 南通 226000)

〔摘要〕 目的/意义 探讨医护科研人员用户画像构建方法和科研能力提升策略。方法/过程 根据用户画像理论提出医护科研人员用户画像的构建模型；使用 Python 语言、ETL 和 Echarts 工具对构建用户画像所需的多源异构数据进行采集、预处理、存储和挖掘，提取标签进行用户画像可视化并进行实证研究。结果/结论 医护科研人员可分为具备科研潜质的护理人员、科研新手、科研提升型群体和科研领域资深专家，根据不同类型特征提出医护科研人员科研能力提升策略。

〔关键词〕 多源数据；医护科研人员；用户画像；科研能力提升

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.07.005

Study on Strategies to Enhance the Scientific Research Capabilities of Medical Personnel under the Fusion of Multi-source Data

YANG Xiaowen¹, JI Hanzhen²

¹ Library of Nanjing Medical University, Nanjing 211166, China; ² Nantong Third People's Hospital, Nantong Third Hospital Affiliated to Nantong University, Nantong 226000, China

〔Abstract〕 **Purpose/Significance** To explore methods for constructing user portraits of medical and nursing research personnel and strategies for enhancing their research capabilities. **Method/Process** Based on user portrait theory, a model for constructing user portraits of medical and nursing research personnel is proposed. Python, ETL and Echarts tools are used to collect, pre-process, store and mine multi-source heterogeneous data required for constructing user portraits, labels are extracted for user portrait visualization and the empirical research is conducted. **Result/Conclusion** Medical and nursing research personnel can be categorized into groups with research potential, research novices, research improvement groups, and senior experts in the research field. Strategies for enhancing the research capabilities of medical and nursing research personnel are proposed based on the characteristics of these different types.

〔Keywords〕 multi-source data; medical research personnel; user portrait; research capability enhancement

〔修回日期〕 2023-11-18

〔作者简介〕 杨晓雯，硕士，副研究馆员，发表论文 25 篇；通信作者：季汉珍，研究馆员。

〔基金项目〕 教育部人文社会科学研究项目（项目编号：22YJC870019）；江苏高校哲学社会科学研究项目（项目编号：2022SJYB0303）；南京医科大学党的二十大精神研究专项（项目编号：XZ221X20230805）。

1 引言

《“健康中国 2030”规划纲要》^[1]对医疗保健体系的现代化、信息化和智慧化提出明确目标，强调大数据、人工智能等现代信息技术在健康领域的深度融合与应用，数据驱动的决策和创新越来越受到重视。人口老龄化加剧以及公共卫生事件频发，对医护科研人员提出更高要求。促进其科研能力、临床能力提升显得尤为迫切。本研究从多源数据融合视角出发，以医护科研人员相关数据为切入点，汇聚与融合异构数据中的共同特征，绘制用户画像，精细化揭示医护科研群体的科研特征和科研行为，探索科研能力提升的有效途径，以期为应对新环境的医疗需求提供新思路和解决方案。

2 研究现状

2.1 国内

国内医学院校图书馆^[2]、医院图书馆^[3]或医院科教科^[4]是研究医护科研人员信息行为的主体，研究方法主要包括问卷调查和专家访谈，研究对象为医生、护士、青年医生^[5-6]、各地区医护人员^[7-9]等。研究范围主要包括医护群体的科研现状、科研困境和科研需求等，从政策、经费、科研环境、科研辅导和文献资源保障角度提出改善医护群体科研能力的对策。既往对医生群体用户画像的研究^[10-12]目的在于构建医生医疗技术擅长领域画像并向患者推荐，对其自身科研兴趣、需求以及医护科研能力提升关注较少。

2.2 国外

国外关于医护科研人员知识服务相关研究主要集中于用户对资源的访问及获取、各类信息行为等方面，例如在资源获取方面，Alving B E^[13]通过文献综述得知医护科研人员最关注的是专业数据库全文获取的便捷性；在需求发现方面，Medina M L 等^[14]提出关注医生科研需求的同时也应关注护理人

员的科研需求；在查询行为方面，Touchette J C^[15]通过对研究对象资源类型偏好、信息搜寻行为方式调查，提出医学图书馆可以为医护科研人员提供学术支持。

2.3 存在不足

医护科研人员知识服务研究多采用问卷调查和专家访谈的方法，存在样本数量较少，研究范围较窄等不足，难以全面深入地探讨医护科研人员知识服务的各个方面。为了弥补以上不足，本研究以医护科研人员科研现状数据为基础，通过对相关数据的采集、预处理、挖掘和可视化等呈现其科研行为和需求。

3 基于多源动态数据的医护科研人员用户画像模型构建

多源数据由不同用户生成，以多种格式描述同一主题^[16]。医护科研人员多源数据全面记录其人口统计、科研属性、成果、合作和行为等信息。根据用户画像构建原则，提出包括数据准备、处理、可视化以及应用服务的医护科研人员用户画像构建流程，见图 1。

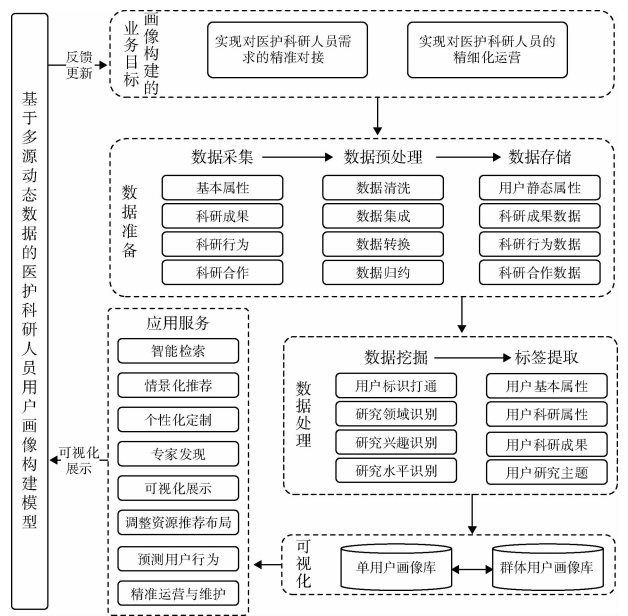


图 1 医护科研人员用户画像构建模型

3.1 数据准备阶段

3.1.1 数据采集 在遵守数据爬取相关法律法规和保护用户隐私的前提下，采用 Python 语言编写爬

虫程序采集网站中与医护科研人员相关的多源数据，通过文献检索与数据库导出采集科研人员的科研成果及科研行为数据，见表 1。

表 1 医护科研人员用户数据库

数据库种类	数据来源	主要内容	数据形式	采集方式
医护科研人员基本属性	医院官网、百度百科、个人主页、好大夫在线、名医汇、微医等包含医生人口统计信息的在线健康社区	姓名、性别、年龄、地域、学历、专业、工作单位、职位、职务、工作经历、头衔	半结构化或非结构化的文本型数据、结构化的日期型数据	爬虫爬取、人工采集
科研成果、科研行为	基金网站、PubMed、Web of Science、SinoMed、中国知网、万方数据、维普资讯	科研基金、学术论文、学术专著、专利成果、导师信息、学术交流信息、合作关系、引证关系	结构化或非结构化的文本型数据	爬虫爬取、人工采集、数据库导出

3.1.2 数据预处理和存储 为处理数据类型不一、噪音、缺失和冗余等问题，利用抽取、转换和加载（extract – transform – load，ETL）工具进行数据清洗、集成、转换和归约。应用 Kettle 工具规整多源异构数据，并以适当格式进行存储。

3.2 数据处理阶段

3.2.1 数据挖掘 利用 K – means 聚类，通过余弦相似度计算属性相似度，隐含狄利克雷分布（latent Dirichlet allocation，LDA）模型推断主题，以及命名实体识别技术提取特定实体。

3.2.2 医护科研人员研究领域识别 运用词频 – 逆向文件频率（term frequency – inverse document frequency，TF – IDF）提取文献摘要特征，应用 K – means 聚类，构建作者 – 关键词/摘要 – 标签三维模型。关键词缺失时，通过文本挖掘算法对摘要进行特征提取以获得所需信息。

3.2.3 医护科研人员科研能力识别 评估医护科研人员的研究能力时，关键指标包括基金课题、学术论文、专著出版和专利申请情况。识别体现研究水平的数据进行分析，如论文发表量、核心期刊发文情况、影响因子、被引次数、基金级别与数量、专著和专利数量。评价研究水平时，筛选出至少发表一篇论文的研究人员，基于其发表论文的平均数设定 4 个研究水平区间：起步、一般、较高和顶尖学术水平。其他指标也用相似方法进行分析和分类。

3.2.4 标签体系构建 根据医护科研人员信息库中所采集的数据及类型，将医护科研人员用户画像标签体系分为 4 个维度，见图 2。

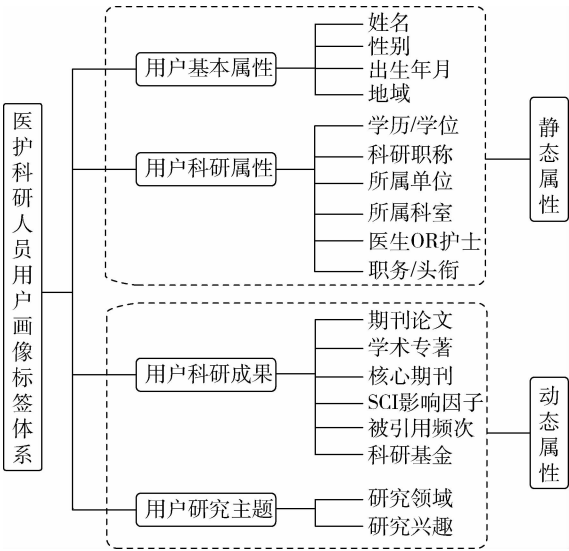


图 2 医护科研人员用户画像标签体系

定义一个四元组作为医护科研人员的向量空间表示：User = { Demographics，Research Properties，Achievement，Status}。（1）基本属性和科研属性标签提取。提出 Demographics = {Name，Sex，Birth，Regional} 表示用户的基本属性模型，Research Properties = {Degree，ProfessionalTitle，institution，Department，Identity，Title} 表示用户科研属性模型。用户基本属性和科研属性是用户的静态属性。（2）科研

成果标签提取。科研成果标签向量模型为 Achievement = {Journal, book, High - quality journal, IF, Citations, Fund}。用户科研成果是用户的动态属性,其中用户科研成果标签主要呈现医护科研人员的科研水平。(3) 研究主题标签提取。研究主题标签向量模型为 Status = {Field, Interest, Resource}。医护科研人员论文成果中的关键词是对文章内容的高度凝练和概括,可以使用关键词作为研究主题的主要表征。

4 应用研究

4.1 数据采集

为了验证医护科研人员用户画像模型的可行性,通过现有数据构建江苏地区肿瘤学科、肝胆外科、心血管内科的医护科研人员画像。人口基本属性和科研属性通过医院官网、百度百科和好大夫在线等获得。共采集 796 名医护科研人员数据,包括肿瘤学科 265 人、肝胆外科 291 人、心血管内科 240 人。科研成果和研究主题数据通过中国知网、PubMed 和 Web of Science 核心合集数据库获取,文献类型选择期刊文献。申请基金情况通过国家自然科学基金网、国家社会科学基金项目数据库以及发表学术论文所标注的基金资助情况获取,最终采集期刊论文信息 39 812 条(中文 23 891 条,外文 15 921 条);基金项目 157 个,其中国家自然科学基金项目 153 个,国家社会科学基金项目 4 个。

4.2 数据处理

不同数据来源导致所采集数据类型各异,使用基于 ETL 技术的 Kettle 工具将多源异构数据进行转换,

实现数据的重组和同构。针对数据缺失问题,通过百度百科、咨询本人或其他渠道补全信息;删除冗余数据,保证数据的唯一性和准确性;通过人名消歧等方法处理同名异义问题^[17]。使用命名实体识别方法对医护科研人员基本简介文本形式数据进行转换。将预处理后数据存入对应数据仓储中。例如,某心血管内科医生向量空间显示如下:

Demographics: {Name: “张明”, Sex: “男”, Birth: “1980 - 01”, Regional: “南京地区”}

Research Properties: {Degree: “博士”, ProfessionalTitle: “副教授”, Institution: “XX 医院”, Department: “心血管内科”, Identity: “医生”, Title: “科主任”}

Achievement: {Journal: 20, Book: 1, High - quality journal: 5, IF: 3.5, Citations: 100, Fund: 3}

Status: {Field: “心血管内科”, Interest: “心肌梗死”, Resource: “PubMed”}

4.3 医护科研人员特征分析及标签提取

医护科研人员的基本属性、科研属性和科研成果标签可直接采用数据仓储中所采集的数据。研究主题标签可通过科研成果中的关键词字段表示,部分关键词标注不全或者数据缺失的文献,可采用 LDA 主题模型进行文本挖掘,找出合适的关键词进行标签赋值。每个主题下词语的概率分布(仅列出部分有代表性的词语),见表 2。根据概率分布推断:主题 A 可能与“肿瘤免疫疗法”有关,主题 B 可能与“心脏病再灌注”有关,主题 C 可能与“肝癌”或“胆囊肝胆外科”有关。查看每个文档的主题分布,将具有最高概率的主题分配给该文档。

表 2 部分词语 LDA 主题文本挖掘

主题	肿瘤	免疫疗法	心脏病	再灌注	肝癌	分子标记物	胆囊	肝胆外科
主题 A	0.70	0.20	0.01	0.02	0.02	0.01	0.01	0.01
主题 B	0.01	0.01	0.60	0.20	0.01	0.01	0.02	0.01
主题 C	0.02	0.01	0.01	0.01	0.60	0.20	0.10	0.20

统计医护科研人员论文成果时,需将中英文关键词统一语种且人工修正之后再行统计,将出现频次

较多的关键词作为科研人员研究领域标签,并利用 SPSS 进行聚类分析和相似度计算。K - means 聚类选

择 4 作为最佳聚类数。随着迭代次数增加，聚类中心不断变化并逐渐收敛，直到获得最终聚类中心和不同聚类包含的个案数，聚类 1 为 152 人，聚类 2 为 325 人，聚类 3 为 223 人，聚类 4 为 96 人。

4. 4 医护科研人员用户画像可视化

采用可视化工具 Echarts，导入所得标签，根据权重设置标签大小。根据用户画像提示，可分为具备科研潜质的护理人员用户画像、科研新手用户画像、科研提升型群体用户画像和科研领域资深专家用户画像，见图 3-6。



图 3 具备科研潜质的护理人员用户画像



图 4 科研新手用户画像



图 5 科研提升型群体用户画像



图 6 科研领域资深专家用户画像

5 医护科研人员特征及科研能力提升策略

5. 1 具备科研潜质的护理人员

该群体包括护士和初级护师，具备临床经验但缺少科研实践。大多未发表科研论文，对科研稍有兴趣并扮演辅助角色。其希望提高科研能力，有意愿参与更多科研项目，但受临床工作压力和科研资源有限等因素影响。可通过以下对策提升该群体的科研能力。一是定制科研培训计划，医院和学术机构共同开发培训方案，强化研究方法论及跨学科数据技能培训。二是建立跨学科合作平台，促进专家与医护科研人员之

间的知识交流, 激发创新并指导医护人员熟练掌握从设计临床试验到科研成果转化的完整科研流程。三是科研资源优化与扶持, 医疗和政府机构应关注科研资源合理分配, 确保研究资金、时间和设施共享, 并通过配备专业导师和举办研讨会增强科研成果转化指导。

5.2 科研新手

该群体包含初涉科研的住院医生及医护人员, 有参与科研项目的经验且已发表论文, 但缺乏领域权威。其需要接受进一步的研究方法、实验设计、数据分析及写作技能培训, 并渴望与资深研究者合作。其面临的问题包括研究方向不明确、缺乏高质量研究指导及资源限制。为增强特定医疗人员科研能力, 可实施以下策略。一是实施定制化培训, 为有潜力的医护人员和科研新手提供包括数据分析、疾病预防知识在内的跨学科研究培训。培训注重研究规划、数据处理、临床实践等实际操作, 并通过导师指导和实验室资源共享帮助其积累科研经验。二是鼓励参与创新项目, 支持新入科研者参与创新性研究, 尤其是与公共健康相关的项目。为其提供从研究设计到资金申请的全方位指导, 鼓励其与资深研究人员或跨学科团队合作, 以快速提升科研实力。三是构建知识共享平台, 建立以老龄化研究为中心的知识共享和协作网络平台, 通过研讨会、工作坊和在线讲座等形式帮助科研人员提升技能、拓宽视野。

5.3 科研提升型群体

该群体为拥有博士学位的资深医学研究者, 已发表权威论文及主持科研项目, 渴望更深层次研究与跨学科合作以增强影响力。面对高质量数据需求、激烈竞争及资金挑战, 其追求持续创新并保持前瞻性。针对该群体可实施以下策略。一是推进跨学科研究, 针对精准医疗等关键议题, 鼓励其启动或参与跨学科项目, 并为其提供必要资源, 如资金和先进研究工具。二是强化国际合作, 通过研讨会和学者交流项目等方式实现全球交流与合作, 提高医学研究的国际影响力。三是提升实用性与政策影

响, 引导其研究侧重实用性, 确保成果转化为医疗政策或实践指南, 并通过建立与决策者的联系, 推动研究在政策层面的应用。

5.4 科研领域资深专家

该群体主要是资深医学教授和科研专家。其在专业领域内已获得广泛的研究成果, 包括高引用学术论文等。此类专家致力于引领学术前沿, 培养新一代研究人员, 并对政策和公共卫生决策产生影响。面临的挑战包括时间管理、维持研究团队的资金与士气, 以及高期望下的持续产出压力。针对该群体提升科研能力的策略一是促进终身学习, 定期组织高级研讨会和讲习班, 确保其掌握最新科研动态和技术。二是加强合作, 由资深专家引导年轻科研人员形成跨学科团队, 促进知识传承与创新。三是优化管理, 通过时间管理工作坊、有效会议技巧等培训, 提升其时间和资源管理能力。提供项目管理和资金筹集支持, 减少行政负担, 使专家能够专注于研究工作。

6 结语

本研究根据用户画像理论提出医护科研人员用户画像构建模型, 并通过实证研究, 将不同类型医护科研人员进行分类和描述。在理论上为医护科研人员用户画像构建提供一种多源数据融合和处理方法, 强调跨学科数据分析和可视化工具应用, 为医学领域用户画像研究提供新思路和方法。在实践上为医疗和政府机构提供有价值的数据分析工具, 有助于更好地理解医护科研人员的特征和需求。本研究尚存在一定局限性: 数据完整性和准确性受到数据源限制, 可能存在一定噪音和缺失; 用户画像建立和标签提取过程依赖于算法和模型, 可能存在一定主观性和误差。后续将开展更多实证研究, 验证提出模型和策略的有效性和可行性。

作者贡献: 杨晓雯负责论文撰写; 季汉珍负责研究设计与论文修订。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 “健康中国 2030”规划纲要 [EB/OL]. [2023-11-04]. https://www.gov.cn/zhengce/2016-10/25/content_5124174.htm.
- 2 丁思瑀, 马路. 附属医院临床医生信息搜寻行为调查——以北京市医学院校附属医院或教学医院为例 [J]. 医学教育管理, 2021, 7 (3): 311-318.
- 3 秦方, 侯胜超, 胡鸿. 医院图书馆用户信息利用行为研究与精准服务对策 [J]. 中华医学图书情报杂志, 2020, 29 (9): 66-70.
- 4 李妹, 陈海燕, 崔若琛. 医学图书馆联盟共享体系建设与成效 [J]. 医学信息学杂志, 2022, 43 (7): 89-93.
- 5 王涵, 边策. 医院微信公众号信息服务研究分析——以北京市为例 [J]. 甘肃科技, 2023, 39 (3): 42-46.
- 6 方芳, 应峻, 陈怡, 等. 复旦大学医学图书馆资源共建共享联盟的实践探索 [J]. 中华医学图书情报杂志, 2017, 26 (6): 30-32.
- 7 郭振宇. 高等医学院校及其附属医院数字文献资源共建共享实践与思考 [J]. 医学信息学杂志, 2016, 37 (10): 76-79.
- 8 杨晓雯, 陈勇, 刘烜贞. “健康江苏”背景下医学高校图书馆知识服务与创新案例 [C]. 北京: 中国图书馆学会年会, 2020.
- 9 王丽丽, 刘贺. 医院图书馆用户信息行为影响因素模型构建及信息服务策略研究 [J]. 中国医院, 2023, 27 (5): 101-104.
- 10 蒋知义, 李巧, 邢思佳. 在线健康社区信息服务质量评价指标体系构建及实证研究 [J]. 情报探索, 2021 (4): 29-36.
- 11 唐晓波, 高和璇. 基于特征分析和标签提取的医生画像构建研究 [J]. 情报科学, 2020, 38 (5): 3-10.
- 12 袁润, 王琦. 学术博客用户画像模型构建与实证——以科学网博客为例 [J]. 图书情报工作, 2019, 63 (22): 13-20.
- 13 ALVING B E, CHRISTENSEN J B, THRYSSØE L. Hospital nurses' information retrieval behaviours in relation to evidence based nursing: a literature review [J]. Health information & libraries journal, 2018, 35 (1): 3-23.
- 14 MEDINA M L, MEDINA M G, MERINO L A. Scientific research as academic mission of the public university hospitals [J]. Revista cubana de salud pública, 2015, 41 (1): 139-146.
- 15 TOUCHETTE J C, KALLMES K M. The value of external research and writing experts to physicians, hospitals, and the scientific community [J]. Journal of hospital management and health policy, 2021 (5): 10.
- 16 莫君兰, 窦永香, 开庆. 基于多源异构数据的科研团队画像的构建 [J]. 情报理论与实践, 2020, 43 (9): 100-106.
- 17 范晓玉. 基于多源科技管理数据的重大项目团队成员推荐研究 [D]. 西安: 西安电子科技大学, 2018.

(上接第 25 页)

- 5 陈朋, 王建生, 马林茂, 等. 国家环境与健康信息元数据标准研究 [J]. 环境与健康杂志, 2013, 30 (9): 821-825.
- 6 文庭孝, 刘晓英. 计量学学科发展谱系研究 [J]. 情报科学, 2022, 40 (6): 12-18, 43.
- 7 Office for National Statistics. Developing the health index for England [EB/OL]. [2023-07-30]. <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/articles/developingthehealthindexforengland/2015to2018>.
- 8 2030 年“健康中国”什么样? ——详解未来中国健康指数 [J]. 中医健康养生, 2016 (12): 2.
- 9 BRODMAN K, ERDMAN A J, LORGE I, et al. The Cornell medical index - health questionnaire VI. The relation of patients' complaints to age, sex, race, and education [J]. Journal of gerontology, 1953, 8 (3): 339-342.
- 10 MURNAGHAN J H. Health indicators and information systems for the year 2000 [J]. Annual review public health, 1981 (2): 299-361.
- 11 程羽, 袁萌, 孙增坤, 等. 不同人群健康指数调查研究 [J]. 河南中医, 2017, 37 (2): 259-262.
- 12 张新平. 部分发达国家卫生制度和健康指数比较分析 [J]. 国外医学 (社会医学分册), 1995 (4): 150-153.
- 13 楚洁, 田雪莹, 姜莹莹, 等. 山东省心血管疾病防控现状评价: 基于中国心血管健康指数研究 [J]. 中国公共卫生, 2021, 37 (6): 954-959.
- 14 高虹. 中国人个人健康评价指标体系研究 [D]. 武汉: 华中科技大学, 2011.
- 15 国民心理健康评估发展中心. 中国心理健康量表 [EB/OL]. [2023-07-30]. <https://cmhr.psych.cn/MentalHealthScale.html>.
- 16 胡蕴绮, 周兰妹. 中文版健康行为能力自评量表的信效度研究 [J]. 中华护理杂志, 2012, 47 (3): 261-262.
- 17 仲来福. 卫生学 [M]. 北京: 人民卫生出版社, 2013.
- 18 朱高培, 孙汉辰, 岳增勇, 等. 山东省健康城市指数构建及其应用 [J]. 中华疾病控制杂志, 2022, 26 (10): 1199-1204.
- 19 唐贤兴, 马婷. 中国健康促进中的协同治理: 结构、政策与过程 [J]. 社会科学, 2019 (8): 3-15.
- 20 平卫伟. 健康测量与评价 [M]. 北京: 科学出版社, 2019.