

基于乳腺癌专病库的知识图谱构建研究^{*}

靳淑雁¹ 王 爽² 黄 琼² 邱五七³ 林怿昊¹

(¹ 深圳市妇幼保健院 深圳 518028 ² 深圳市卫生健康发展研究与数据管理中心 深圳 518028

³ 中国医学科学院/北京协和医学院医学信息研究所 北京 100020)

〔摘要〕 目的/意义 创建乳腺癌知识图谱的本体模型, 实现乳腺癌知识的结构化表达和可视化展示。方法/过程 对乳腺癌知识进行抽取和表示, 利用 Protégé 软件创建乳腺癌知识图谱本体模型。模型包含患者、患者基本信息、检查、诊断、检查实体属性值 5 类实体, 以及具有、检查、诊断、属于等 10 余种关系。结果/结论 乳腺癌知识的图形化展示有助于提高公众对乳腺癌的关注和了解, 对乳腺癌的预防与健康管理具有积极作用。

〔关键词〕 专病库; 乳腺癌; 本体构建; 医学知识图谱; 人工智能

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.12.011

Study on the Construction of Knowledge Graph Based on Breast Cancer Specialized Disease Database

JIN Shuyan¹, WANG Shuang², HUANG Qiong², QIU Wuqi³, LIN Yihao¹

¹ Shenzhen Maternity & Child Healthcare Hospital, Shenzhen 518028, China; ² Shenzhen Health Development Research and Data Management Center, Shenzhen 518028, China; ³ Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔Abstract〕 **Purpose/Significance** To construct the ontology model of breast cancer knowledge graph, and to realize the structured expression and visual display of breast cancer knowledge. **Method/Process** Breast cancer knowledge is extracted and represented, and the ontology model of breast cancer knowledge graph is constructed by Protégé software. The model includes 5 types of entities including patient, basic information of patient, inspection, diagnosis and attribute value of inspection, and more than 10 kinds of relationships such as having, examining, diagnosing, belonging and so on. **Result/Conclusion** The graphical display of breast cancer knowledge helps to improve the public's attention and understanding of breast cancer, and has a positive effect on the prevention and health management of breast cancer.

〔Keywords〕 specialized disease database; breast cancer; ontology construction; medical knowledge graph; artificial intelligence

1 引言

知识图谱 (knowledge graph) 概念于 2012 年由谷歌提出, 基本组成单位为三元组, 通过实体之间

的关系构成网状知识结构^[1-3]。知识图谱可以分为两个层次: 概念层和实例层。实例层由事实组成, 知识以事实为单位进行存储; 概念层构建在实例层之上, 通过本体库来规范实例层的事实表达^[4-7]。近年来医学数据激增, 海量患者数据可供知识发

〔修回日期〕 2023-11-21

〔作者简介〕 靳淑雁, 主任医师, 发表论文 30 余篇; 通信作者: 林怿昊。

〔基金项目〕 广东省医学科学技术研究基金项目 (项目编号: C2021106); 广东省深圳市基础研究专项 (自然科学基金) 面上项目 (项目编号: JCYJ20210324134405016); 广东省深圳市卫生经济学会科研基金项目 (项目编号: 202331)。

现、管理、共享及应用,通过构建医学知识图谱,可以对这些知识进行管理、应用及共享^[8-9]。深圳市医防融合乳腺癌专病库是一个全面规范的专病库,系统整合了乳腺癌患者信息和诊疗信息。本文基于深圳市医防融合乳腺癌专病库,结合知识图谱技术基本原理,对专病库中的数据进行知识图谱本体构建,通过 Protégé 软件,创建包括乳腺癌患者基本信息、检查、诊断等内容的本体模型,并对其中部分概念进行可视化展示。通过将知识图谱与专病库结合,提出乳腺癌知识图谱的本体构建方法,有助于提高公众对乳腺癌的关注和了解,对乳腺癌的预防与健康管理具有积极作用。

2 医学知识图谱研究现状

刘俊丽等^[10]以万方医学网收录的两年内乙肝领域国内核心期刊文献为数据源,采用文本挖掘技术,进行数据清洗、实体抽取、构建共词矩阵并绘制该领域知识图谱,证实此方法适用于生物医学领域的知识探索与发现。赵雪娇^[11]对妇产科教材中的医学知识进行抽取和表示,将妇产科知识存储成结

构化知识图谱。侯梦薇等^[12]对知识图谱在医疗领域应用进行了综述,将医学知识表示、医学知识抽取、医学知识融合和医学知识推理所运用的关键技术作为重点。袁凯琦等^[13]将医学知识的表示、抽取、融合、推理以及质量评估作为构建技术的 5 个主要组成部分。崔洁等^[14]尝试采用电子病历(electronic medical record, EMR)数据构建疾病知识图谱。在对 EMR 数据应用方面,谷歌领衔加利福尼亚大学等国际一流医学院提出一种基于快速医疗互操作性资源格式的患者完整、原始电子健康档案记录表示,结合 3 个深度学习神经网络模型结构,证明该深度学习方法能够准确地预测来自多个中心的医疗事件^[15]。Ma F 等^[16]基于 3 个真实电子健康档案数据集将离散的先验医学知识与预测技术结合起来。由于互联网数据发展高度动态化,翟社平等^[17]为了改善知识图谱补全更新实时性欠佳的问题,运用贝叶斯网络概率推理图模型与本体推理规则,定量分析知识图谱节点间的依赖程度。

3 医防融合乳腺癌专病库

3.1 专病库架构(图 1)

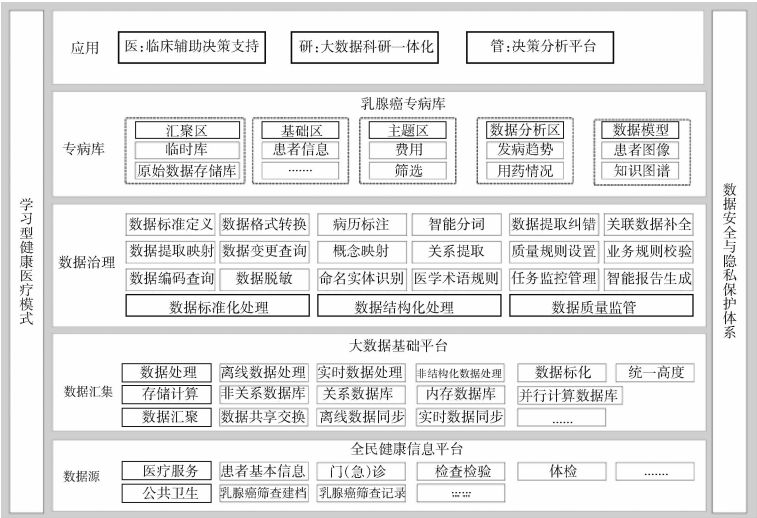


图 1 专病库架构

深圳市医防融合乳腺癌专病库(以下简称专病库)整合全市乳腺癌相关患者及其诊疗信息,接入特定疾病患者诊疗和治疗数据,根据该疾病的发生发展规律、临床诊疗方案,开展全面系统的数据清

洗、标准化、归一化、后结构化处理,形成统一规范的乳腺癌专病库。专病库的数据范围为 2017 年 1 月—2023 年 4 月,数据总容量为 5.8 GB,数据总量约 1 577 万条,数据范围包括门(急)诊病历数据、

就诊记录、检查检验信息、手术记录、入院记录、治疗记录、治疗用药、过敏记录、死亡记录、人口学信息数据及病案首页数据。专病库建设以大数据平台为架构基础，以学习型健康医疗模式为指导，以深圳市全员人口健康信息平台为数据源，按照最小数据集标准开展乳腺癌病种的专病库建设。

3.2 专病库建设流程与数据的后结构化

3.2.1 专病库建设流程 依据《深圳市重大疾病防治基本数据编制标准项目》《医防融合基本数据

集标准编制项目》建立乳腺癌单病种库数据模型。定期或实时采集乳腺癌医疗数据，并存储在集中的数据仓库中；对数据进行清洗，并对采集汇聚的数据进行加工处理，通过去重、转类型、脱敏等过程，生成易于临床分析利用的数据^[18]。通过流程和规则库的规范操作，设置数据转换、比对、关联、融合等过程，对重复数据、无效值、错误数据、残缺数据及缺失值等进行删除、纠错、补全处理，形成一个完善的处理机制^[19]，为下一个学习周期提供可靠、真实的数据，见图2。

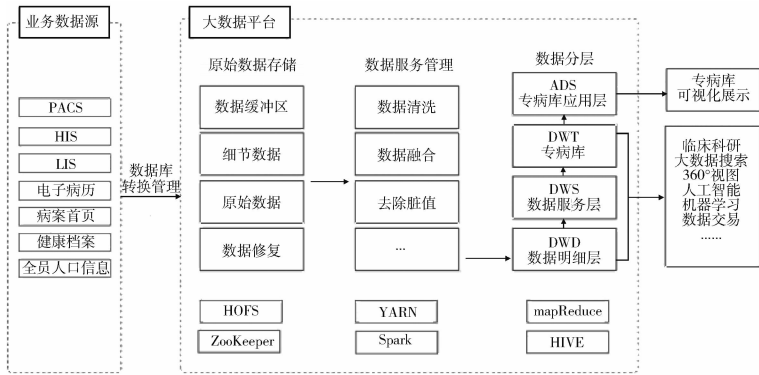


图2 专病库建设流程

3.2.2 专病库后结构化流程 通过自然语言处理技术，将临床数据中的非结构化文本数据，经过自动分词、词性分析、句法分析和语义分析等工作，转换成可识别的结构化数据，实现后结构化处理，为临床科研和辅助诊疗提供语义理解层面的数据支撑。通过自然语言处理、机器学习、本体构建和规则提取等技术融合，识别半结构化/非结构化文本中的医学命名实体（如疾病、症状、手术、药物等），提取实体的属性，将数据结构化，见图3。

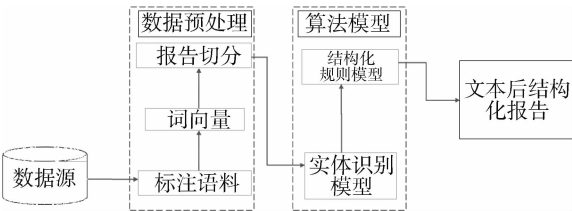


图3 专病库结构化流程

4 乳腺癌知识图谱构建方法

基于乳腺癌专病库，通过自顶向下的方式定义本体与数据模式的概念层，进而添加实体、构建实例层。该方法包括概念层设计和实例层设计，通过知识建模、知识抽取、知识融合、知识存储与加工、知识推理与应用等关键技术^[20]，实现医学知识图谱构建。

4.1 概念层设计

以乳腺癌专病库结构化数据为基础，结合《2022版<中国临床肿瘤学会乳腺癌诊疗指南>更新解读》^[21]，构建乳腺癌知识图谱的概念层。概念层包括患者、患者基本信息、检查、诊断、检查实体属性值5个概念级别；包含10余种关系类别，如

具有、检查、诊断、属于、属性、住院诊断名称、医学影像检查报告、手术操作名称、检查分类、检查项目名称等。对已构建的乳腺癌实体类进行数据属性定义，见表 1。

表 1 乳腺癌实体类数据属性

实体类	属性
患者	医疗卡号等
患者基本信息	年龄、性别、地区等
检查	检查申请单号、检查项目代码、检查部位、检查目的、申请单状态、临床信息（病史资料）等
诊断	诊断流水号、诊断编码、诊断编码类型、诊断名称、主要诊断标志、疑似诊断标志等
检查实体属性值	检查类型代码、检查类型名称、检查设备仪器型号、检查仪器号、检查申请日期、检查科室编码、检查科室名称、检查部位、检查结果描述、检查结果（定性）、检查结果参考值（定性）、检查定量结果等

4.2 实例层设计

在医学知识图谱中，医学信息抽取的核心任务在于实体识别、实体消歧及关系抽取^[22]。实例层设计属于信息抽取范畴，即从无语义信息的病历记录中抽取与概念层相匹配的医疗事实知识^[14]。与其他领域相比，医学领域关系抽取的标准更加严格，且医学实体间的关系也更加复杂。专病库中既有门（急）诊病历信息、就诊信息、检验检查信息等结构化数据，又有出院小结等半结构化数据，此外还有超声文本报告等非结构化数据。

4.2.1 提取乳腺癌医学实体 首先要从专病库数据中找到描述乳腺癌医学实体的医学术语集合。根据乳腺癌专病库中实际的病历数据，概念层包括患者、患者基本信息、检查、检查实体属性值、诊断 5 类实体。从患者性别、年龄和地区等字段，提取相应字段值作患者基本信息的属性值。检查实体包含病理检查、超声检查和电子计算机断层扫描（computed tomography，CT）检查等不同类型的子实体，在经过结构化处理后的检查报告中

可以提取相关检查值作为检查实体属性值；不同检查对应的诊断包含病理诊断、超声诊断和 CT 诊断等子实体。

4.2.2 提取实体间的关系类型 经过医学实体抽取，乳腺癌患者的病历记录可形成 5 类医学实体，通过与概念层的属性相匹配，这些实体可以作为概念属性的属性值^[16]。由上述概念层的 5 类实体，结合实体之间的关系，可以得出以下关系：患者与患者基本信息之间的具有关系，患者与检查之间的检查关系，患者与诊断之间的诊断关系；检查与超声检查、CT 检查、病理检查之间，诊断与超声诊断、CT 诊断、病理诊断之间，均为实例关系；检查项与检查报告之间的隶属关系等。提取出实例层的实体与实体间的关系后，形成专病库中乳腺癌知识图谱三元组关系，见表 2。

表 2 知识图谱的三元组

实体	关系	实体
患者	具有	患者基本信息
患者	检查	检查
患者	诊断	诊断
检查	实例	超声、CT、病理检查
诊断	实例	超声、CT、病理诊断
超声检查报告的检查项	属于	超声检查报告
病理检查报告的检查项	属于	病理检查报告
检查报告的检查值	属性	检查报告的检查项
超声诊断结论	属性	超声诊断
病理诊断结论	属性	病理诊断

5 乳腺癌知识图谱的设计

基于专病库，结合知识图谱技术的基本原理，按照自顶向下的模式设计和构建乳腺癌知识图谱的本体。乳腺癌专病库中某位乳腺癌患者的结构化数据记录，见表 3。乳腺癌患者数据患者、患者基本信息（性别、年龄和地区等）、检查（病理检查、超声检查和 CT 检查等）和诊断（病理诊断、超声诊断和 CT 诊断等）等实体信息^[14]。

表 3 乳腺癌专病库中某位患者的数据（部分）

实体类	实体	实体属性值
基本信息	性别	女
	年龄	28
	地区	深圳
超声检查	位置	右
	方位	12 点
	大小	11 mm × 11 mm × 6 mm
	形状	不规则形
	边界	不清
	边缘	毛刺状
	内部回声	极低回声
	分布	不均匀
	回声	点状强回声伴声影
	后方回声	无变化
	血流信号	条状
	血管走行	弯曲
超声诊断	血流速度	较快
	诊断结论	BIRADS4c

Protégé 软件^[23]是美国斯坦福大学的本体编辑和知识获取软件，基于 Java 语言开发，屏蔽具体的本体描述语言，提供本体概念类、关系、属性以及实例创建，用户可以在概念层上创建领域的本体模型^[24]。本文使用 Protégé 软件构建患者、患者基本信息、检查、诊断、检查实体属性值等 5 类实体，并通过定义实体对象属性及实体数据属性，实现对乳腺癌知识的建模^[25]。将结构化数据中的各项数值对应到概念层中的实体与关系中，通过前期构建好的三元组表，得到乳腺癌知识图谱的实例层。乳腺癌相关概念可以进行图形可视化展示。用户可以交互式浏览乳腺癌相关概念，并选择其中一个概念来进行下一步的信息挖掘或检索。

6 结语

医学知识图谱能够解决当前医疗数据丰富庞

杂而知识管理困难的问题。本文基于乳腺癌专病库，提出知识图谱本体构建方法，并详细阐述概念层和实例层的设计流程，同时利用 Protégé 工具，创建乳腺癌知识图谱的本体模型，实现乳腺癌知识的结构化表达和可视化展示。乳腺癌知识图谱的构建为医学知识图谱用作辅助诊断与智能问答奠定了数据基础，也为其他病种库的数据管理提供借鉴。

参考文献

1 范媛媛, 李忠民. 中文医学知识图谱研究及应用进展 [J]. 计算机科学与探索, 2022, 16 (10): 2219 – 2233.

2 李绪辉, 阎思, 陈沐坤, 等. 面向真实世界的知识挖掘与知识图谱补全研究 (一): 真实世界数据与知识图谱概述 [J]. 医学新知, 2023, 33 (2): 130 – 135.

3 修晓蕾, 吴思竹, 崔佳伟, 等. 医学知识图谱构建研究进展 [J]. 中华医学图书情报杂志, 2018, 27 (10): 33 – 39.

4 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45 (4): 589 – 606.

5 WONG W, LIU W, BENNAMOUN M. Ontology learning from text: a look back and into the future [J]. ACM computing surveys (CSUR), 2012, 44 (4): 1 – 36.

6 STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: principles and methods [J]. Data & knowledge engineering, 1998, 25 (1 – 2): 161 – 197.

7 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53 (3): 582 – 600.

8 ZHANG Y, CHEN K, SLOAN S A, et al. An RNA – sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex [J]. Journal of neuroscience, 2014, 34 (36): 11929 – 11947.

9 WANG M, LIU M, LIU J, et al. Safe medicine recommendation via medical knowledge graph embedding [EB/OL]. [2023 – 09 – 23]. <https://arxiv.org/abs/1710.05980v2>.

10 刘俊丽, 张秀梅, 蒋勇青. 基于文本挖掘的乙型肝炎相关文献知识图谱分析 [J]. 医学信息学杂志, 2014, 35 (1): 48 – 53.

11 赵雪娇. 妇产科知识图谱构建研究与实现 [J]. 中国数

- 字医学, 2019, 14 (1): 3-5.
- 12 侯梦薇, 卫荣, 陆亮, 等. 知识图谱研究综述及其在医疗领域的应用 [J]. 计算机研究与发展, 2018, 55 (12): 2587-2599.
- 13 袁凯琦, 邓扬, 陈道源, 等. 医学知识图谱构建技术与研究进展 [J]. 计算机应用研究, 2018, 35 (7): 1929-1936.
- 14 崔洁, 陈德华, 乐嘉锦. 基于 EMR 的乳腺肿瘤知识图谱构建研究 [J]. 计算机应用与软件, 2017, 34 (12): 122-126.
- 15 RAJKOMAR A, OREN E, CHEN K, et al. Scalable and accurate deep learning with electronic health records [J]. NPJ digital medicine, 2018, 1 (1): 18.
- 16 MA F, GAO J, SUO Q, et al. Risk prediction on electronic health records with prior medical knowledge [C]. London: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.
- 17 翟社平, 郭琳, 高山, 等. 一种采用贝叶斯推理的知识图谱补全方法 [J]. 小型微型计算机系统, 2018, 39 (5): 995-999.
- 18 王飞, 黄艺璠, 汪鹏. 基于多模态数据的肺癌专病库建设研究 [J]. 中国数字医学, 2021, 16 (12): 85-88, 104.
- 19 吴骋, 王志勇, 徐蕾, 等. 基于人工智能的电子病历数据质量控制 [J]. 解放军医院管理杂志, 2021, 28 (2): 134-135, 168.
- 20 封颖, 叶祥, 彭敬, 等. 企业信息系统审计知识图谱的构建方法与应用 [J]. 审计月刊, 2022, (12): 5.
- 21 孙正魁, 江泽飞. 2022 版《中国临床肿瘤学会乳腺癌诊疗指南》更新解读 [J]. 中国肿瘤外科杂志, 2022, 14 (3): 212-218.
- 22 匡泽民, 李健铨, 邓楠. 信息抽取在构建医学知识图谱中的应用及进展 [J]. 医学信息学杂志, 2021, 42 (1): 7.
- 23 PROTÉGÉ A. A free, open source ontology editor and framework for building intelligent systems [EB/OL]. [2023-09-23]. <https://protege.stanford.edu/>.
- 24 刘志, 郝克俊. 基于 Protégé 的人工影响天气术语本体知识库设计与实现 [J]. 中国科技术语, 2019, 21 (6): 17.
- 25 杨美洁, 熊相超. 基于爬虫技术和电子病历的糖尿病知识图谱的构建 [J]. 中国数字医学, 2020, 15 (2): 6-8.

(上接第 59 页)

- 11 朱丽丽, 张国. 常态化疫情防控背景下医院开展医疗工作实践及思考 [J]. 中国医院, 2021, 25 (4): 84-86.
- 12 梁旭健, 曾锐峰, 邓志玲, 等. 基于微信平台的医院药学服务实践: 现状与启示 [J]. 中国医院药学杂志, 2022, 42 (22): 2421-2424, 2429.
- 13 王一夫. 2020 中国互联网医疗诊后: 改善资源配置刻不容缓 [J]. 互联网经济, 2020 (12): 90-95.
- 14 任安霖, 潘思, 邓海骏, 等. 基于医护到家平台的护士参与“互联网+护理服务”现状调查 [J]. 护理学杂志, 2020, 35 (5): 56-59.
- 15 任宇飞, 庾兵兵, 杨冲, 等. 线上线下一体化互联网医院云平台建设实践 [J]. 中华医院管理杂志, 2020, 36 (10): 837-840.
- 16 张恺, 赵移畛. 河南省 49 家二、三级中医医院信息化建设现状及问题分析 [J]. 世界中西医结合杂志, 2023, 18 (3): 609-614.
- 17 郭敏, 周晓英, 宋丹, 等. “互联网+”时代的我国医院微信信息服务研究 [J]. 图书与情报, 2015 (4): 19-25.
- 18 刘玉霞. 数字经济时代下的智慧医院发展分析 [J]. 医学信息学杂志, 2022, 43 (2): 52-56.
- 19 林伟良, 张健, 楼东升. 基层医疗机构应用全门诊可视化智慧导诊系统的实践 [J]. 中国农村卫生事业管理, 2018, 38 (12): 1549-1551.
- 20 计虹, 王梦莹. 医院智慧服务综合体系架构设计与应用 [J]. 中国卫生信息管理杂志, 2022, 19 (1): 23-27.
- 21 汪瑾, 冷锴, 陆慧. “互联网+”视域下智慧医疗服务模式创新研究 [J]. 南京医科大学学报 (社会科学版), 2020, 20 (1): 84-87.