·医学信息研究·

基于循证医学和电子病历数据的通用医学知识图谱构建*

吴 欢[1,2] 车贺宾[1,2] 王万玲[1,2] 陈媛媛[1,2] 李重勋[3] 张 超[3] 庄 严[1,2] 何昆仑[1,2]

([1] 中国人民解放军总医院医学创新研究部 北京 100853
[2] 医疗大数据应用技术国家工程研究中心 北京 100853
[3] 北京嘉和海森健康科技有限公司 北京 100085)

〔摘要〕 **目的/意义** 构建涵盖循证医学知识和电子病历数据的通用医学知识图谱，以提升图谱的应用效能。**方法/过程** 梳理多源异构数据情况，融合国内外知名知识图谱，设计图谱 schema。利用 RoBERTa 预训练模型进行词嵌入，从医学文献、网络文献、教科书、医学数据库和电子病历等数据源中提取命名实体和关系，采用基于规则的 SWIQA 框架和基于随机抽样的人工审核策略评价图谱质量。**结果/结论** 共确定 128 个本体和 1 108 种关系，并以三元组形式存储于数据库中。经评估，图谱语义准确性达 93.8%。所构建的通用医学知识图谱不仅涵盖循证医学知识，还包括临床真实世界产生的专家经验，为医学人工智能应用的进一步发展提供了有力支撑。

〔关键词〕 循证医学；电子病历；医学知识图谱
〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2025.02.004

Construction of a General Medical Knowledge Graph Based on Evidence-based Medicine and Electronic Medical Record Data

WU Huan[1,2], *CHE Hebin*[1,2], *WANG Wanling*[1,2], *CHEN Yuanyuan*[1,2], *LI Zhongxun*[3], *ZHANG Chao*[3], *ZHUANG Yan*[1,2], *HE Kunlun*[1,2]
[1]*Medical Innovation Research Department of Chinese PLA General Hospital, Beijing 100853, China*; [2]*National Engineering Research Center of Medical Big Data Application Technology, Beijing 100853, China*; [3]*Goodwill Hessian Health Technology Co. Ltd., Beijing 100085, China*

〔**Abstract**〕 **Purpose/Significance** To construct a general medical knowledge graph covering evidence-based medical knowledge and electronic medical record (EMR) data, so as to improve the application effect of the graph. **Method/Process** The multi-source heterogeneous data are sorted out, the well-known knowledge graphs at home and abroad are integrated, the schema of the graphs is designed. The word embedding of RoBERTa pre-trained model is used to extract entities and relationships from medical literatures, network literatures, textbooks, medical databases and EMRs. The rule-based SWIQA framework and the manual audit strategy based on random sampling are used to evaluate the quality of the graph. **Result/Conclusion** A total of 128 ontologies and 1 108 relationships are identified and stored in the database in the form of triples. The semantic accuracy of the atlas is 93.8% by evaluation. The general medical knowl-

〔修回日期〕 2024-08-05
〔作者简介〕 吴欢，工程师，发表论文 20 篇；通信作者：何昆仑，博士，教授。
〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目（项目编号：2021ZD0140406）；北京市自然科学基金-海淀原始创新联合基金资助项目（项目编号：L222006）。

edge graph not only covers evidence – based medical knowledge, but also includes expert experience generated in the clinical real world, which can provide support for the further development of medical AI applications.

〔**Keywords**〕 evidence – based medicine; electronic medical record (EMR); medical knowledge graph

1 引言

知识图谱研究最早可追溯到 20 世纪 50 年代的语义网络模型，随着互联网和人工智能技术的发展，相关研究和应用在各行各业迅速扩展。2012 年谷歌公司正式提出“知识图谱”概念，其本质是一个基于语义网的知识库，主要由节点和边组成。医学是知识图谱应用最广泛的领域之一，已成为国内外人工智能研究热点。

目前国内较为成熟的医学知识图谱包括中医药学知识图谱[1]、医学百科知识图谱[2]、CMeKG[3]等。其主要基于临床指南、专家共识和文献书籍等循证医学知识构建而成，尽管这些知识具有较高的可靠性和准确性，但仍存在知识更新缓慢、无法完全覆盖真实应用场景等问题[4]。电子病历（electronic medical record, EMR）覆盖个人健康医疗数据并不断积累，被视为辅助医疗的最核心资源。因此，研究者开始利用自然语言处理和机器学习技术，通过补充真实世界的 EMR 数据提升知识图谱的全面性。2018 年何霆等[5]分析当时医疗知识图谱研究的热点及其演变，认为基于 EMR 的医疗知识图谱研究尚处于起步阶段。2020 年胡佳慧等[6]以临床文本数据为主要研究对象，从知识发现生命周期、文本处理流程和关键技术等方面，研究临床文本数据处理与知识发现的方法。2022 年 Zhao G 等[7]提出基于 EMR 数据提取医疗实体的方法，并定义了一个名为 MLEE 的计算体系结构，用于提取具有“对象 – 属性”依赖关系的对象级实体。2023 年 Murali L 等[8]总结基于 EMR 数据构建医学知识图谱的局限性和难点，指出当前研究主要局限于通用技术，较少关注知识图谱构建中真实数据源的开发。而基于 EMR 的知识图谱构建面临着数据复杂性和维数高、知识融合不足、知识图谱动态更新等挑战。

本研究旨在结合循证医学和真实世界 EMR，挖掘真实世界 EMR 数据中的知识，以解决医学知识图谱缺乏现实实践经验的问题。首先，采用自顶向下和自底向上相结合的方式，通过定义事件本体，构建基于循证医学知识和 EMR 的通用医疗领域知识图谱 schema。同时，通过查阅国内外医疗术语体系，结合临床数据特征，为各医疗本体制定相应的术语基线。其次，针对多本体间关联数据场景，采用语义关系网络框架定义 SPO（subject – predicate – object）域。最后，通过实体识别、关系抽取等技术，完成基于循证医学文献知识和真实世界 EMR 数据的通用医学知识图谱构建，并采用 SWIQA 框架[9]对知识图谱质量进行了评估。

2 知识来源及关键技术

2.1 知识来源

目前医学知识图谱的构建主要基于医学书籍、临床指南、临床路径、药品说明书等循证医学知识，PubMed 和中国知网文献是知识图谱中科学文献的主要知识来源，开放的医学术语库也是循证医学知识的重要来源，如 UMLS[10]、SNOMED – CT[11]、ICD – 10、DrugBank[12]、NPC（national drug codes）、ICD – 9 – CM – 3 等。EMR 系统中记录了患者基本信息、就诊住院、检查检验和治疗等信息，是真实世界数据的主要来源，包括结构化数据（如检验结果）和自由文本形式的非结构化数据（入院记录、手术记录、护理记录等）。

2.2 技术路径

采用自顶向下和自底向上相结合的知识图谱构建方法，见图 1。在本体顶层模式构建中，遵循本体构建标准采用自顶向下方法，包括本体整体设计、详细设计、本体表示和本体评价等步骤，设计

符合医院实际情况的概念模型。在实体学习中，采用自底向上的方法，基于临床知识和真实世界临床数据进行知识抽取、融合，得到多模态通用医学知识图谱。

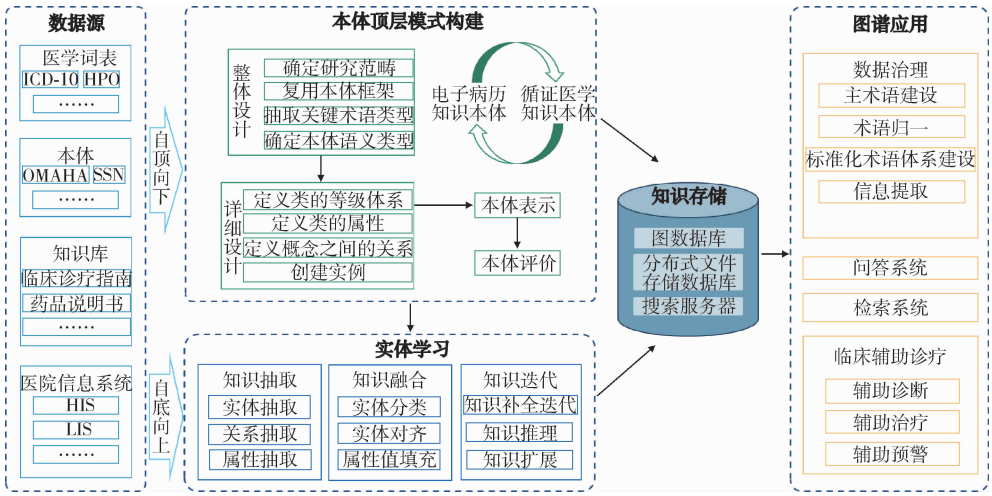

图 1 知识图谱构建技术路径

2.3 本体 schema 设计

根据国内外医学知识图谱概念层定义的通用范式，规范定义不同本体在医学范畴的内涵与外延，同时为满足医院通用型知识图谱下游任务需求，考虑各本体的特定描述、属性和关系，最终讨论得到包含事件本体在内的 128 个本体，1 108 种关系。部分通用本体类型，见表 1。

表 1 通用医学知识图谱本体介绍（部分）

序号	本体类型	本体描述	数据标准或子事件
1	疾病本体	表示在一定的病因作用下，机体内稳态调节紊乱而导致的生命活动障碍	全国版 RC020 – ICD – 10 诊断编码
2	药品本体	表示由生产厂家生产的药物制剂，名称通常包括物质、强度和剂型等。包括中成药、中草药、西药	其中西药采用药物 ATC 编码
……	……	……	……
124	就诊过程事件	患者的一次入院事件	门诊事件、住院事件、急诊事件
125	检测事件	指在医学诊疗过程中发生的对患者进行所有检测事件的全称	医学检验事件、医学检查事件
126	诊断事件	指在医学诊疗过程中发生的对患者进行所有诊断事件的全称	会诊事件、问诊事件
127	治疗或预防事件	指在医学诊疗过程中发生的对患者进行所有治疗或预防相关事件的全称	开药事件、治疗事件、手术事件
128	临床不良事件	指不在疾病范围内的，发生的可导致不良反应的事件；指正常剂量的药物用于预防、诊断、治疗疾病或调节生理机能时出现的有害的和与用药目的无关的反应	药品不良反应事件

2.4 关键技术

本体 schema 设计完成后，开发数据模型支撑医疗数据处理，实现对各类数据的知识抽取。数据包括结构化和半结构化数据，结构化数据采用规则映射，半结构化数据要先用信息抽取技术（如命名实体识别、关系抽取、属性抽取等）进行知识抽取，再将知识映射到数据模型。数据来源包括电子病历数据和医学文献数据，二者抽取方式大致相同，主要区别在于数据文本的分布差异导致的关系抽取论元时触发词不同。电子病历中文本主要是针对单一患者的病情描述，关系触发词多为针对发病与治疗

过程的客观描述，如：行……术、服用……药物、胸痛两天等；而文献数据主要是循证医学知识，关系触发词更加客观，如：患病率、证据支持、临床表现等。基于以上情况，构建命名实体识别和关系抽取方法架构。

2.4.1 命名实体识别 知识图谱主要由头实体、尾实体、关系、属性及属性值构成。实体不仅是知识图谱中最基本的组成元素，其抽取任务也是构建知识图谱的关键步骤。命名实体识别的质量与知识图谱的质量息息相关。采用基于 RoBERTa 的预训练模型、双向长短期记忆网络（bidirectional long short－term memory，BiLSTM）和条件随机场（conditional random field，CRF）相结合的 RoBERTa＋BiLSTM＋CRF 框架进行命名实体识别，见图 2，RoBERTa[13]的上下文敏感特性可通过预训练语言模型捕捉复杂的输入序列特征，BiLSTM 是递归神经网络的变体，其双向特性有助于理解文本中的前后文关系，CRF 具备序列优化能力，确保输出的标签序列达到全局最优[14]。

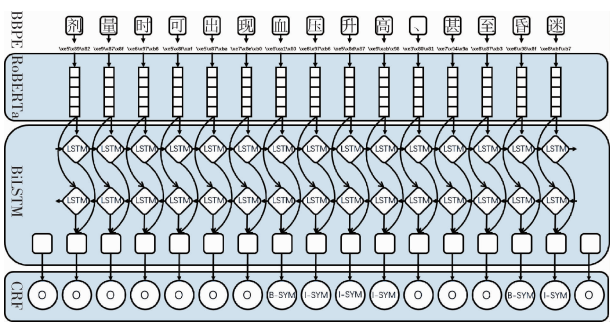

图 2 命名实体识别技术架构

2.4.2 关系抽取 关系抽取分为两个阶段，先利用 RoBERTa 预训练模型和特征工程相结合的框架抽取关系词；再利用 RoBERTa 预训练模型、条件层归一化（conditional layer normalization）框架以及前一阶段抽取的关系词预测主体和客体。第 1 阶段的关系词抽取模型架构结合 RoBERTa 预训练模型的深度学习能力和经典特征工程技术。首先，将标注数据中的所有触发词作为触发词抽取的先验特征。其次，将句子中与标注触发词相匹配的文本标注出来，与 RoBERTa 的输出在其他特征（嵌入）层进行拼接。最后，得到用于判断触发词起始和结束位置索引的两个二分类向量，见图 3。RoBERTa 预训练模型是一个高效的实体识别系统，但其输出可能无法全面覆盖所有实体类型和复杂的实体形式。因此引入特征工程，通过整合词性标注、实体边界标识及局部文本特征等信息，强化模型对于不同实体关系的识别和分类能力。此外，识别潜在的关系触发词也是本阶段的关键任务，这些词是表征实体间关系的关键词或短语。因此设计特定组件并训练其从文本中提取可能的关系触发词，包括动词、名词短语及其他可能指示实体关系的词汇元素。

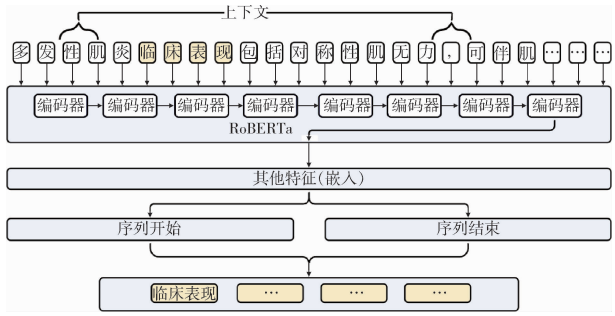

图 3 关系词抽取模型架构

第 2 阶段进行关系词实体识别，技术路线图，见图 4。首先，采用 RoBERTa－Conditional Layer Normalization 作为关系提取器，将原始文本和触发词在文本中的位置作为输入。其次，在 RoBERTa 模型对原始文本进行编码时，利用第 1 阶段的触发词在文本中的位置作为条件，将 RoBERTa 的归一化层替换为条件归一化层，使模型可以更敏感地捕捉到细微的上下文差异，这些差异通常是区分不同实体关系的关键。最后，使用文本中所有词到触发词的相对距离作为额外信息与编码向量进行拼接，得到用于判断主体和客体起始和结束位置索引的 4 个二分类向量，可分别用于判断针对此触发词关系的主体和客体。

2.5 知识融合与更新

知识抽取完成后，不同数据来源的知识存在知识冗余、格式不统一和关系不明确等问题，采用框架匹配和实体对齐两种方式分别对本体和实体进行消歧、加工和整合。首先，利用框架匹配技术对本

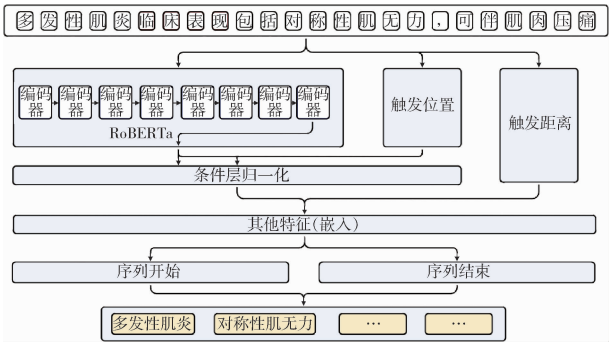

图 4 关系词实体识别模型

体或知识图谱体系进行融合，通过计算本体或知识图谱之间的相似性，逐个分析各本体属性或关系的范围，对相同定义下的属性关系内容进行排序，删除传递信息量相同的信息，组合优化传递信息量相似的信息，补充到相应的属性关系下，实现本体层之间的融合。其次，结合文本语义和上下文信息，利用实体对齐技术实现实例层之间的融合，实现相似、近义实例的统一。抽取的知识经过框架匹配、实体对齐之后以三元组的形式存储在数据库中，用于支撑上层数据应用。

知识来源不断更新，知识库及模型需要循环迭代，知识库迭代路径及方法，见图 5。首先，利用训练好的任务模型，分别进行命名实体识别、关系抽取。其次，将抽取的内容以自动化和人工审核相结合的方式进行数据过滤处理，并依据 schema 结构进行实体建立与关系链接，更新到已有知识库中。最后，将知识库内容提炼成满足模型训练要求的数据格式以进行模型再训练，选取效果更好的任务模型进行替换，提高内容抽取的准确性。

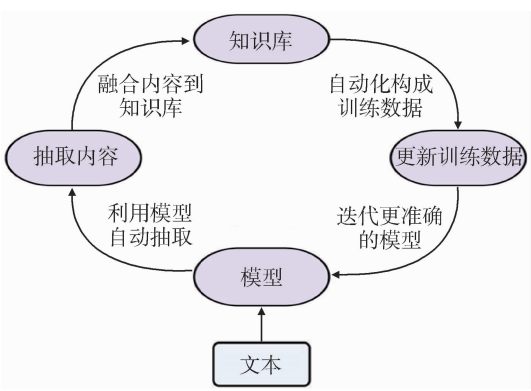

图 5 知识库迭代路径

3 通用医学知识图谱构建

研究中循证医学资料包括临床指南和专家共识 7 000 余份，中国知网和 PubMed 文献约 91 万份，EMR 数据来源于中国人民解放军总医院约 4 万例心脑血管疾病患者。利用上述技术手段对文献资料、电子病历中的非结构化和半结构化文本数据进行命名实体识别、实体关系识别和抽取，并对同一实体的不同名称进行实体对齐与知识更新迭代。基于融合循证医学知识和电子病历数据的通用医学知识图谱 schema 框架，构建一整套包含常见医疗实体的通用型医学知识图谱。知识图谱疾病本体中一个实体的具体实例，见图 6。

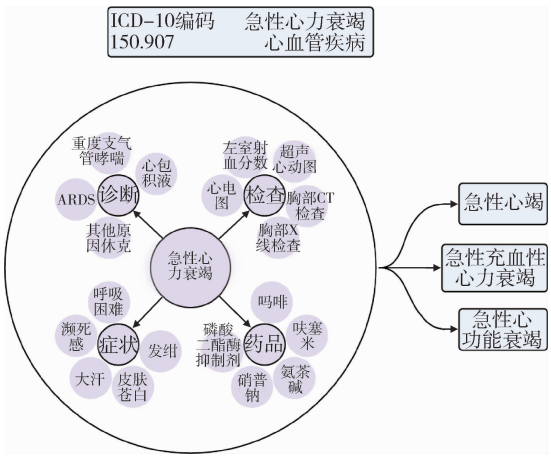

图 6 知识图谱疾病本体实例

以急性心力衰竭为例，在实体属性方面，可归类为心血管疾病，编码字段记录该实体在 ICD－10 中的编码信息，通过实体对齐，将急性心竭、急性充血性心力衰竭、急性心功能衰竭列为急性心力衰竭的同义词。在实体间关系方面，急性心力衰竭在事实上可与诊断、检查、症状、药品等本体中实体产生事实性关联，可基于关系抽取方法将抽取到的关系及实体以三元组形式存储。如急性心力衰竭的常见症状包括呼吸困难、大汗、发绀、皮肤苍白、濒死感等，用三元组表示为＜急性心力衰竭，临床表现，呼吸困难＞等。每个三元组均为一个独立的关系，前文图 6 展示了急性心力衰竭在 4 种不同关系类型下各自独立的三元组关系。

本研究对 schema 中定义的 128 个本体对应的实

体进行了关系抽取、实体对齐等操作，将知识库中30 余万实体进行实体融合与关系抽取，最终得到约200 万个三元组对，得到知识库数据情况，见表 2。

表 2 知识库数据统计（部分）（个）

本体名称	实体数量	层级关系数量	同义词关系数量	上位术语
临床科室	173	140	0	54
疾病	31 704	23 180	19 764	22 890
检查	4 215	3 873	8 833	2 242
检验	14 644	2 307	11 151	768
药品	154 877	0	0	0
非手术治疗	3 433	885	1 152	671
手术	26 361	13 952	8 202	8 419
症状	93 831	4 800	9 009	17 998

4 知识图谱质量评估

知识图谱质量考察对象主要包括概念、实体、属性 3 类个体知识对象，以及概念之间的关系、概念与实体之间的关系、实体之间的关系 3 类关系知识对象。知识图谱质量评估的内容主要涵盖语义准确性、语法准确性和完整性 3 个方面，SWIQA 框架利用一些手工设计的规则来判断三元组的正确性，但因规则设计存在不全面性，该框架主要侧重于语法正确性和完整性的评估，语义准确性评估方面存在不足。因此另外采用基于随机抽样的人工评价方式进行语义准确性评估。

4.1 SWIQA 框架

评估知识图谱语法准确性最简单有效的方式是根据语法规则对当前知识图谱中需要关注的语法要点进行定义。本研究主要对合法值和异常值进行语法准确性评估，见表 3。完整性通过重复数据清理和数据引用完整性验证来评估。相关语法规则设计，见表 4。

表 3 基于 SWIQA 框架的知识图谱语法准确性规则（部分）

语法规则名	定义	举例
句法规则	定义字符属性的类型或文字值的模式	值必须只包含字母
合法值规则	对某一属性的允许值进行明确定义	性别的值域只能是“男”“女”“未知”
合法值域规则	对数值属性允许的值范围进行明确定义	年龄必须是大于 0 的值
关系型字段归属规则	实体中关系型字段的所属本体和指向本体均应在知识图谱的本体列表中	症状本体的字段“症状发生位”归属于人体形态与结构本体。确保上述字段的主体和客体都属于本体列表

表 4 基于 SWIQA 框架的知识图谱完整性规则（部分）

语法规则名	定义	举例
未删除规则	检查所有关联 ID，在对应本体中是否存在字段“是否删除实体”且值为“否”	“胸痛”的发生部位“胸口”。确保“胸口”的“是否删除实体”字段值为“否”
被引用规则	假如当前实体有被其他实体引用，即其他实体中存在指向该实体的关系，则称该实体被引用	实体“后背”在“心绞痛”的“放射部位”字段中，即“后背”为被引用实体
孤立实体规则	假如当前实体既没有被其他实体引用，也没有引用其他实体，则称该实体为孤立实体。这种实体越少越好	该实体的 ID 没有被其他实体引用，同时该实体中的所有关系型字段值也均为空，即满足孤立实体规则

为评估数据引用完整性，对构建的知识图谱进行统计分析，见表 5。定义“孤立实体”为该实体既没有被其他实体引用，也没有引用其他实体。通过计算得到，“被引占比”最高和最低的比例分别为 96.32% 和 9.70%，而“孤立占比”最高值为 13.04%，最低值为 0.01%，证明所构建知识图谱内部的相互引用关系具有较高完整性。

表 5 知识图谱统计分析（部分）

本体名称	实体总量(个)	被引实体(个)	被引占比(%)	孤立实体(个)	孤立占比(%)
组织	7 305	5 841	79.96	1	0.01
药品	154 877	68 309	44.11	642	0.41
手术	26 361	9 281	35.21	192	0.73
部位	2 187	962	43.99	34	1.55
检查	14 211	1 706	12.00	226	1.59
成分	3 345	3 222	96.32	65	1.94
临床科室	336	86	25.60	12	3.57
疾病	61 414	16 181	26.35	2 207	3.59
检验	14 644	1 420	9.70	778	5.31
人	23	13	56.52	3	13.04

4.2 语义准确性评估

首先，基于 SWIQA 框架确保绝大多数实体都有相应的关系连接，确保关系型字段的主体和客体均满足合法值规则。其次，开发可视化工具，用于知识图谱数据的可视化审查，便于发现节点群之间的聚集效应，以突出数据核心节点，明确语义核查重点。最后，设计随机抽取策略，对实体关系完整性检查和可视化审查无法完全涵盖的情况进行人工审核，确保知识图谱数据的准确性和完整性。评估所构建知识图谱的语义准确性为 93.8%。

5 结语

循证医学知识图谱具有质量高、可信度高的优点，但 EMR 数据记录着患者在院的整个活动过程，是医疗信息的核心数据，涵盖患者病情相关信息。将循证医学知识与 EMR 数据相结合，可为医疗活动提供更加全面的参考依据。然而，EMR 数据中包含大量自由文本，其中含有重要信息且形式多样、结构复杂，从中准确提取医学信息存在一定难度。知识图谱对此提供了一种新的解决方案，通过构建结构化、一致性和准确性的知识表示方法，确保从医疗记录到知识库构建的每个步骤准确无误，为基于知识图谱的上层应用奠定基础，辅助医生在复杂的临床决策环境中作出更明智的选择。未来将进一步整合机器学习和人工智能技术，探索知识图谱在提升临床治疗效果、增强患者安全性以及优化医疗资源配置等方面的潜在作用。

作者贡献：吴欢负责研究设计、图谱构建、论文撰写；车贺宾、王万玲、陈媛媛负责图谱构建、论文修订；李重勋、张超负责知识抽取、质量评估；庄严、何昆仑负责研究设计、论文指导。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 贾李蓉，刘静，于彤，等．中医药知识图谱构建［J］．医学信息学杂志，2015，36（8）：51－59.

2 刘燕，傅智杰，李姣，等．医学百科知识图谱构建［J］．中华医学图书情报杂志，2018，27（6）：28－34.

3 奥德玛，杨云飞，穗志方，等．中文医学知识图谱 CMeKG 构建初探［J］．中文信息学报，2019，33（10）：1－9.

4 赵悦淑，王军，王蕊，等．中文医学知识图谱研究进展［J］．中国数字医学，2021，16（6）：86－91.

5 何霆，吴雅婷，王华珍，等．基于 EHR 的医疗知识图谱研究与应用综述［J］．哈尔滨工业大学学报，2018，50（11）：137－144.

6 胡佳慧，赵琬清，方安，等．基于医疗大数据的临床文本处理与知识发现方法研究［J］．中国数字医学，2020，15（7）：11－13，88.

7 ZHAO G，GU W，CAI W，et al. MLEE：a method for extracting object－level medical knowledge graph entities from Chinese clinical records［J］. Frontiers in genetics，2022，13（7）：1－12.

8 MURALI L，GOPAKUMAR G，VISWANATHAN D M，et al. Towards electronic health record－based medical knowledge graph construction，completion，and applications：a literature study［J］. Journal of biomedical informatics，2023，143（7）：104403.

9 CHRISTIAN F，HEPP M. SWIQA－a semantic web information quality assessment framework［EB/OL］.［2023－12－26］. https：//aisel.aisnet.org/cgi/viewcontent.cgi?article=1075&context=ecis2011.

10 BODENREIDER O. The unified medical language system（UMLS）：integrating biomedical terminology［J］. Nucleic acids research，2004，32（1）：267－270.

11 DONNELLY K. SNOMED－CT：the advanced terminology and coding system for ehealth［J］. Studies in health technology and informatics，2006，121：279－290.

12 WISHART D S，KNOX C，GUO A C，et al. Drugbank：a comprehensive resource for in silico drug discovery and exploration［J］. Nucleic acids research，2006，34（1）：668－672.

13 LIU Y，OTT M，GOYAL N，et al. Roberta：a robustly optimized bert pretraining approach［EB/OL］.［2023－12－26］. https：//arxiv.org/pdf/1907.11692.

14 AN Y，XIA X，CHEN X，et al. Chinese clinical named entity recognition via multi－head self－attention based BiLSTM－CRF［J］. Artificial intelligence in medicine，2022，（5）：127.