数智医学专题

[专家简介] 何昆仑，解放军总医院医学创新研究部某中心主任，主任医师、教授，主要从事医疗大数据、医学人工智能、智能卫勤装备和系统等研究。担任智能医学工程国家级重点实验室主任，承担重大项目等27项，制定国家行业标准29项，授权发明专利103项，获国家科技进步二等奖1项，省部级一、二等奖5项。担任中国研究型医院学会副会长、北京医师协会副会长等。

医疗大数据平台建设需求、实施路径与成效探讨

吴 欢，车贺宾，乌日力格，王万玲，陈媛媛，何昆仑
解放军总医院医学创新研究部，医疗大数据应用技术国家工程研究中心，解放军总医院医学工程实验室，国家药监局人工智能医疗器械研究与评价重点实验室，北京 100853

摘要：背景 随着医疗信息化的发展，医疗大数据平台成为临床研究资源再分析、利用的关键突破点。然而，医疗数据的多源异构性、数据标准多样性、患者隐私保护要求高等特点增加了数据采集与应用的难度。**目的** 分析医疗大数据平台建设需求，研发自助式全流程数据治理平台及工具，构建多中心医疗大数据平台。**方法** 通过梳理医院数据应用需求，采用模块化、组件化的构建思路进行平台架构设计，提炼出与应用系统相对独立、通用的组件及管理工具，搭建多中心、多源异构医疗大数据平台。**结果** 完成解放军总医院门急诊和住院电子病历数据汇聚治理，研发了全流程、可视化数据治理工具。定义的事件图谱Schema涵盖29个本体类别和128个概念、1 009种关系和3 022种属性，包括临床循证医学知识和临床诊疗、物联网、医学影像等数据。数据治理的一致性和可溯源性达99.99%，知识准确率达到95%以上。构建了贯穿科研全流程、覆盖不同研究类型需求的一站式数据智能检索与科研分析系统和专病库智能分析系统。**结论** 该平台不仅为临床科研人员提供了数据检索与分析系统，还为数据工程师提供了数据治理和平台运维工具，提升了平台的可扩展性和灵活性。

关键词：医学大数据；多源异构；数据治理；模块化；多中心大数据平台
中图分类号：R319；TP311. 13 **文献标志码：**A **文章编号：**2095-5227(2025)02-0119-07
DOI：10.12435/j.issn.2095-5227.24070102
引用本文：吴欢，车贺宾，乌日力格，等.医疗大数据平台建设需求、实施路径与成效探讨［J］.解放军医学院学报，2025，46（2）：119-125.

Research on demand, implementation pathways and effectiveness of medical big data platform

WU Huan, CHE Hebin, WU Rilige, WANG Wanling, CHEN Yuanyuan, HE Kunlun
Medical Innovation Research Department of PLA General Hospital, National Engineering Research Center of Medical Big Data Application Technology, Medical Engineering Laboratory of PLA General Hospital, Key Laboratory of Artificial Intelligence Medical Device Research and Evaluation of the National Medical Products Administration, Beijing 100853, China
Corresponding author: HE Kunlun. Email: kunlunhe@plagh.org

Abstract: Background With the development of medical informatization, medical big data platforms have become an important foundation for clinical research and a key breakthrough point for resource re analysis and utilization. However, the multi-source heterogeneity of medical data, the diversity of data standards, and the high requirements of patient privacy protection increase the difficulty of data acquisition and application. **Objective** To analyze the requirements for establishing a medical big data platform, develop a self-service, full-process data governance platform and tools, and construct a multi-center medical big data platform. **Methods** By reviewing the application needs of hospital data and adopting a modular and component-based design

收稿日期：2024-07-01
基金项目：新一代人工智能国家科技重大专项(2021ZD0140406)
第一作者：吴欢，硕士，工程师。Email: wh2531@126.com
通信作者：何昆仑，博士，主任医师，博士生导师。Email: kunlunhe@plagh.org

approach, the platform architecture was designed. The universally applicable components and management tools that were relatively independent of specific application systems were extracted and utilized to build a multi-center, multi-source heterogeneous medical big data platform. **Results**　The electronic medical record data of the outpatient, emergency and inpatient departments of the Chinese PLA General Hospital had been aggregated and processed. A full-process, visual data governance tool was developed. The defined event schema graph covered 29 ontology categories, 128 concepts, 1 009 relationships and 3 022 attributes, including clinical evidence-based medical knowledge, clinical diagnosis and treatment, Internet of Things, medical imaging and other data. The consistency and traceability of data governance reached 99.99%, and the knowledge accuracy rate was over 95%. A one-stop data intelligent retrieval and scientific research analysis system and a specialized disease database intelligent analysis system covering the entire scientific research process and different research types have been constructed. **Conclusion**　The platform not only provides a data retrieval and analysis system for clinical researchers, but also provides data management and platform operation and maintenance tools for data engineers, which improves the scalability and flexibility of the platform.

Keywords: medical big data; multi-source heterogeneous; data governance; modularization; multi-center big data platform

Cited as: Wu H, Che HB, Wu RLG, et al. Research on demand, implementation pathways and effectiveness of medical big data platform[J]. Acad J Chin PLA Med Sch, 2025, 46(2): 119-125.

医疗数据作为医疗行业重要的信息资源，除具有传统大数据的数量大、种类多、产生速度快等特点之外，还具有复杂性、精确性、隐私性、异构性及封闭性等特点。医疗数据主要包含临床诊疗数据、健康监测数据、管理运营数据和规则知识数据4大类。其中，临床诊疗数据主要包括患者的基本信息、诊断信息、治疗信息、检查检验数据、影像资料等数据资源，涉及结构化数据、半结构化数据、非结构化数据，来源于不同的信息系统，如医院信息系统(hospital information system，HIS)、电子病历系统(electronic medical records，EMR)、实验室信息系统(laboratory information system，LIS)、医学影像存档与通信系统(picture archiving and communication system，PACS)等。数据汇聚、治理、融合难度大，特别是由于医院各类信息系统建设年代不同、厂家产品各异等，存在数据标准不统一、完整度不够、质量参差不齐、安全隐患大等问题，因此智能化的高质量数据治理是提高科研产出、支持医疗数据应用的关键[1-3]。

解放军总医院为提高临床科研数据应用效率，打造医工结合新范式，启动了医疗大数据平台建设规划，组织行业内专家对平台建设需求和实施路径进行了多轮专项论证。本研究总结了平台建设过程中的需求分析、建设路径及应用成效。

1　医疗大数据平台建设需求

1.1　统一的数据标准需求

医院临床数据来自不同的信息系统，由于缺少统一的数据模型，未遵循统一的标准术语规范体系，随着技术发展迭代更新、数据种类不断扩增、应用场景持续拓展，可能存在数据语义不一致、术语标准不统一、编码格式不规范等问题，从而导致无法高效地将各类数据进行深度整合，难以形成高质量的数据资产，很难支撑多中心、大队列的科研协作[4-5]。因此，为开展全周期医疗健康大数据服务，充分发挥医院健康医疗大数据价值，需加强数据标准化建设，实现以患者为中心的汇聚融合。

1.2　个性化的数据治理需求

医疗文本数据包括电子病历数据、检查报告等，涵盖了患者诊疗过程中的重要信息，准确地提取文本中的信息有助于提高临床科研效果。由于各医院和科室对医疗文本的内容结构、书写要求不尽相同，利用单一通用的自然语言处理(natural language processing，NLP)模型和术语体系进行文本数据结构化和标准化处理，效果欠佳。因此，需针对不同文本类型、不同专科专病进行NLP模型训练，通过命名实体识别、语义消歧补全等，提升文本数据提取率和结构化准确率，通过数据归档和应用共享，不断积累术语知识，扩大医学术语识别范围，提高术语归一化效果。

1.3　灵活的变量定义和复用

随着医疗技术的不断发展和临床科研需求的不断变化，平台需要不断升级和完善，以适应新的需求和技术环境。目前针对科室提出的科研数据需求通常有两种处理方式，一种是建设科研平台或专病库系统，用户可以自行检索和申请导出数据，但随着科研需求不断变化，既有的科研平台或专病库无法满足数据要求；另一种则是医院数据库工程师根据科室提供的数据要求进行数据提取，投入成本大、效率低。因此，亟须应用大数据和人工智能技术，增强科研平台敏捷响应、快速适配能力，支撑实现科研数据变量的灵活自

主定义和复用，提升平台的可扩展性和可维护性。

1.4 便捷的检索与分析功能

临床科研不仅要求科研人员有足够的时间整理数据，还要求其具备专业的数据分析能力。随着大数据、人工智能技术的发展，临床科研场景越发丰富，传统的统计分析方法难以满足科研需求，而新技术的发展要求研究者具备一定的统计分析背景和编程能力。此外，医学研究中对患者纳排标准设定通常比较复杂，包含多个纳排条件的灵活组合，通常为年龄、性别、诊断、治疗方式、并发症、检验检查结果、药品等的“或、且、非”逻辑组合，如何找准、找全待研究人群是医学科研的前提，因此平台需具备满足复杂医学检索逻辑的数据检索功能和数据溯源能力，保证研究人群检索的准确性和全面性[6-7]。为了发现隐藏在数据中的规律和价值，平台应具备便捷、准确的数据分析和挖掘功能，帮助科研人员更好地理解和解释数据，快速实现临床问题探索与发现。

1.5 流程化的数据管控能力

临床数据涉及患者的隐私和权益，因此在平台建设和使用过程中，必须严格遵守相关法律法规，确保数据的合法获取和使用，数据不出平台逐步成为数据应用的基本安全要求[8]。传统的临床科研分析系统由于不支持人工智能等算法，需将数据从平台导出到其他软件中处理，导入导出过程中存在数据泄露等安全风险。为此，通过搭建隐私保护与数据安全测试环境，为临床科研人员提供数据、计算、用户隔离的开箱即用分析服务，保证数据的可获取、可重用、可互操作，实现科研结果的可验证和可重现[9]。

2 平台架构及建设路径

2.1 平台架构设计

基于以上建设需求，本研究利用机器学习和深度学习等人工智能技术，聚焦数据治理核心技术研究，设计基于语义本体的通用数据模型和术语标准，制定数据质量和价值评估机制，构建医疗大数据治理体系，形成数据治理工作流和系列工具，研发医疗大数据智能检索与科研分析系统、专病库智能分析系统、临床辅助决策系统等[10]。平台架构设计见图1。

医疗大数据平台整体架构设计包括以下4层。

(1)数据资源层：临床数据主要包括医疗物联网数据、医疗信息系统的结构化数据、非结构化电子病历数据、组学数据等，知识主要来源于图书、指南、共识、文献和患者真实世界临床数据。

(2)数据抽取集成层：通过构建可视化ETL工具，基于统一的数据模型将分散的数据集中存储在临床数据中心。

(3)数据融合治理层：数据融合治理包括数据治理和知识加工，数据治理主要包括数据标准建设、数据处理、数据治理与数据质控，知识加工包括知识抽取、融合与迭代，最终形成涵盖临床数据和医学知识的多源异构医疗大数据平台。

(4)数据应用层：基于构建的医疗大数据平台，开展医疗大数据智能检索与科研分析系统、临床专病数据智能分析系统等应用研发。

2.2 平台建设路径

2.2.1 数据体系和知识体系的构建 为了解决数据治理不透明的问题，平台构建了一种全流程、透明化数据治理系统(图2)，包括制定数据输入规范、优化数据流程和全流程质量预警等。首先，制定数据输入规范是通过明确数据治理目标、责任和流程，包括数据所有权、访问权限、质量标准、分类和保护措施等规定，确保所有数据管理活动都能按照统一的标准和流程进行。其次，优化数据流程是明确数据的转移、抽取、治理和存储方式及数据使用的限制和规范，包括数据访问权限的管理、备份和恢复策略、安全措施等内容，确保数据的合法性、安全性和隐私保护。最后，全流程质量预警是通过PDCA(plan-do-check-action cycle)流程逐步提高数据治理透明度和数据治理质量，包括任务报表、数量报表以及维度报表，确保数据完整性、一致性、准确性和及时性等，实现质量评价维度全流程覆盖。

数据模型是多源异构数据汇聚的基础，有助于划定数据抽取范围，指导数据抽取与治理工作的开展[11-13]。通过对解放军总医院业务系统的梳理，构建了涵盖病案首页、急诊、检查检验等主要业务数据的多源异构数据整合模型(图3)，包括结构化诊疗业务数据54张表，病历文书24张表。

数据标准制定是数据治理的核心问题，也是决定医疗大数据治理能否成功的关键一步。结合国内外成熟的医学术语体系，如医学主题词表(Medical Subject Headings，MeSH)[14]、国际疾病分类与代码(ICD10/ICD9)[15]、观测指标标识符逻辑命名与编码系统(Logical Observation Identifiers Names and Codes，LOINC)[16]、中文一体化医学语

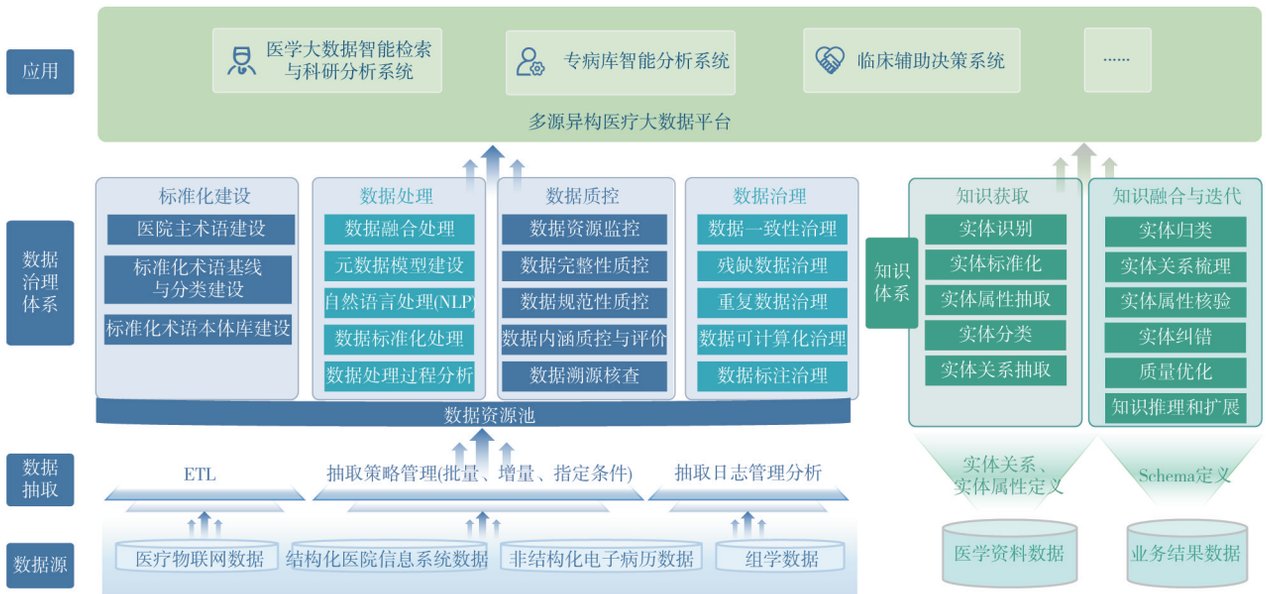

图 1　医学大数据平台架构设计图

Fig. 1　The platform architecture of medical big data

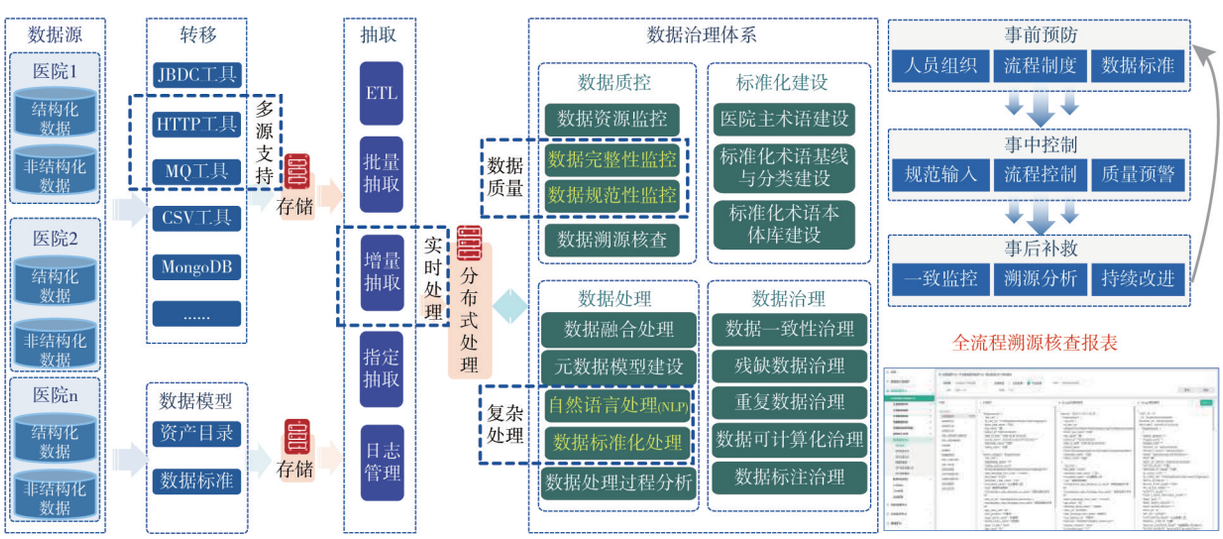

图 2　多源异构数据治理

Fig. 2　The data governance of multi-source heterogeneous data

言系统[17]、中文医学主题词表(Chinese Medical Subject Headings，CMeSH)等，制定平台的术语集。为了医学数据能够以统一的标准形式被各类信息处理方法充分利用，进而形成一套有效的医学知识表示体系，平台建设过程中，在数据治理体系之外，构建了一套知识体系。与常见的医学知识图谱相比，通过抽取出临床诊疗事件，形成了患者维度的事件图谱，实现对临床数据的深加工或二次加工。

2.2.2　智能化数据处理工具的研发　在临床科研平台建设项目中，采用模块化、组件化的构建思路，通过对应用系统中共性问题的归纳，提炼出与应用系统相对独立、通用的组件及管理工具，作为系统统一的底层，解决因数据治理或科研需求不断变化而重复建设的问题。面向数据工程师搭建自然语言处理中心、知识图谱中心、术语中心、变量中心、表单中心等工具(图 4)，以术语、规则、变量、知识生产和加工为基本功能，实现业务数据向结构化、标准化的转换与复用。通过建立标准化数据交换及应用标准，在医疗业务系统与应用系统之间构建中间支撑及服务平台，实现平台灵活、可扩展的开放架构。通过该数据处理工具集，实现非结构化文本数据提取、专病数据变量定义与取数、知识图谱构建等功能，提升数据治理效率。

知识图谱中心为医学知识管理平台，通过三元组构建知识图谱，实现医学术语统一管理。自然语

图 3　多源异构数据整合模型

Fig. 3　The model of multi-source heterogeneous data integration

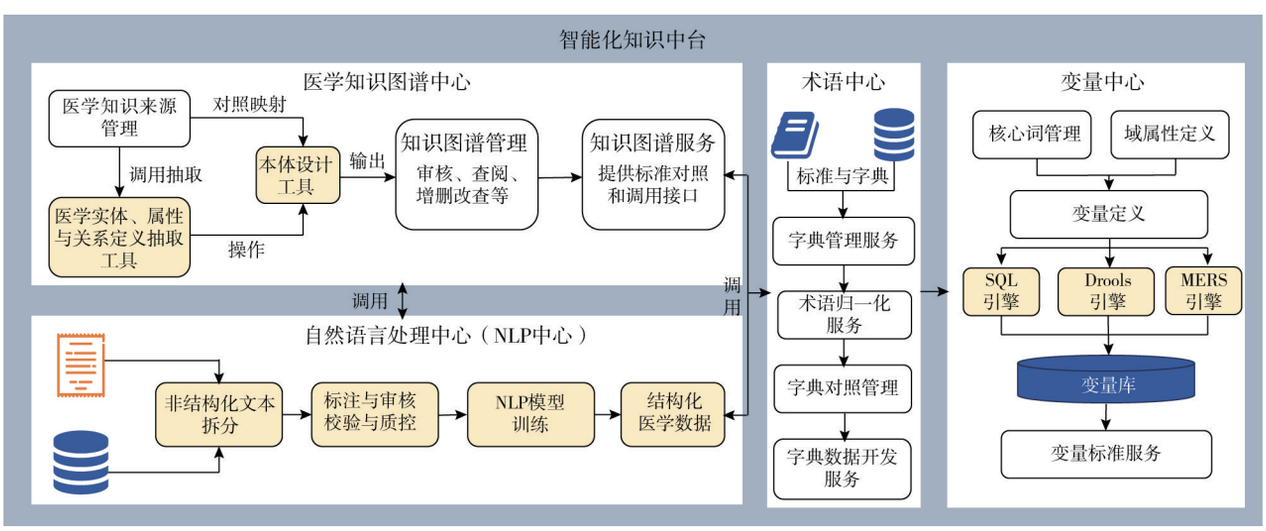

图 4　智能化数据处理工具

Fig. 4　Intelligent data processing tools

言处理中心通过提供NLP模型训练及管理平台，实现基于院内真实诊疗数据训练的模型效果提升。术语中心提供术语对照、归一化服务等，实现数据标准化处理。变量中心支持变量定义、数据生产模型的统一管理，实现复杂医学逻辑数据生产需求。

2.2.3　科研专病一体化系统的研发　面向全院临床科研人员搭建医疗大数据智能分析与专病数据系统，提供人群检索、数据集构建、数据清洗、分析建模、报告生成、数据权限审批等全流程数据应用功能，实现自助式数据应用，减少数据工程师工作量。此外，在大数据平台通用数据治理基础上，通过构建表单中心，进行专病数据库表单配置、模板维护等，实现专科专病数据库的快速构建和数据的二次加工。

系统研发过程中，针对数据分析专业门槛高、统计编辑难等问题，设计了一个内置算法池。将高影响因子期刊中归纳总结出的20余种临床常用统计分析算法纳入算法池中，并在系统中设置算法使用说明、相关案例、智能校验、多模型分析、模型集成分析、参数调整、结果解读等功能，支

持平台内临床分析研究，避免了数据导入导出的安全问题，还降低了学习成本和专业门槛。通过归纳总结的临床常用分析场景，将统计分析场景化，形成统计分析路径，实现统计分析从点到面的突破，大大缩短临床科研周期。

2.2.4　数据安全与应用管理　平台通过数据治理过程中的脱敏措施、应用过程中的权限管控、平台仅内网部署等方法保障数据安全与数据应用之间的平衡。目前，平台通过制定统一的数据治理标准，实现解放军总医院多个医学中心的数据整合与治理；通过不展示隐私字段、隐私字段信息去隐私化等手段进行患者隐私保护；通过平台的内网部署，实现多中心内网环境的统一登录；通过按角色、用户、功能等权限管控模块，实现多中心数据利用权限管控(如可设置当前用户或角色是否支持多中心数据检索)。此外，为了保障数据的应用安全，智能检索与分析平台中通过设置数据导出审批功能，支持根据字段、是否加密等进行数据导出，实现外单位数据合作。

3　建设成效

基于以上设计，现已完成解放军总医院门急诊和住院电子病历数据汇聚治理，研发了全流程、可视化数据治理工具。定义的事件图谱Schema包括了29个本体类别和128个概念、1 009种关系和3 022种属性，涵盖临床循证医学知识和临床诊疗、物联网、医学影像等数据。数据治理的一致性和可溯源性达99.99%，知识准确率达到95%以上。图5展示了数据库中某患者入院事件相关的图谱。

构建了贯穿科研全流程、覆盖不同研究类型需求的一站式数据智能检索与科研分析系统和专病库智能分析系统(图6)，完成627个通用变量和1 814个心脑血管专病变量定义及数据生产。

4　结语

目前，北京协和医院、北京大学第三医院、四川大学华西医院、上海交通大学医学院附属瑞金医院等国内大型三甲医院均陆续完成院级医疗

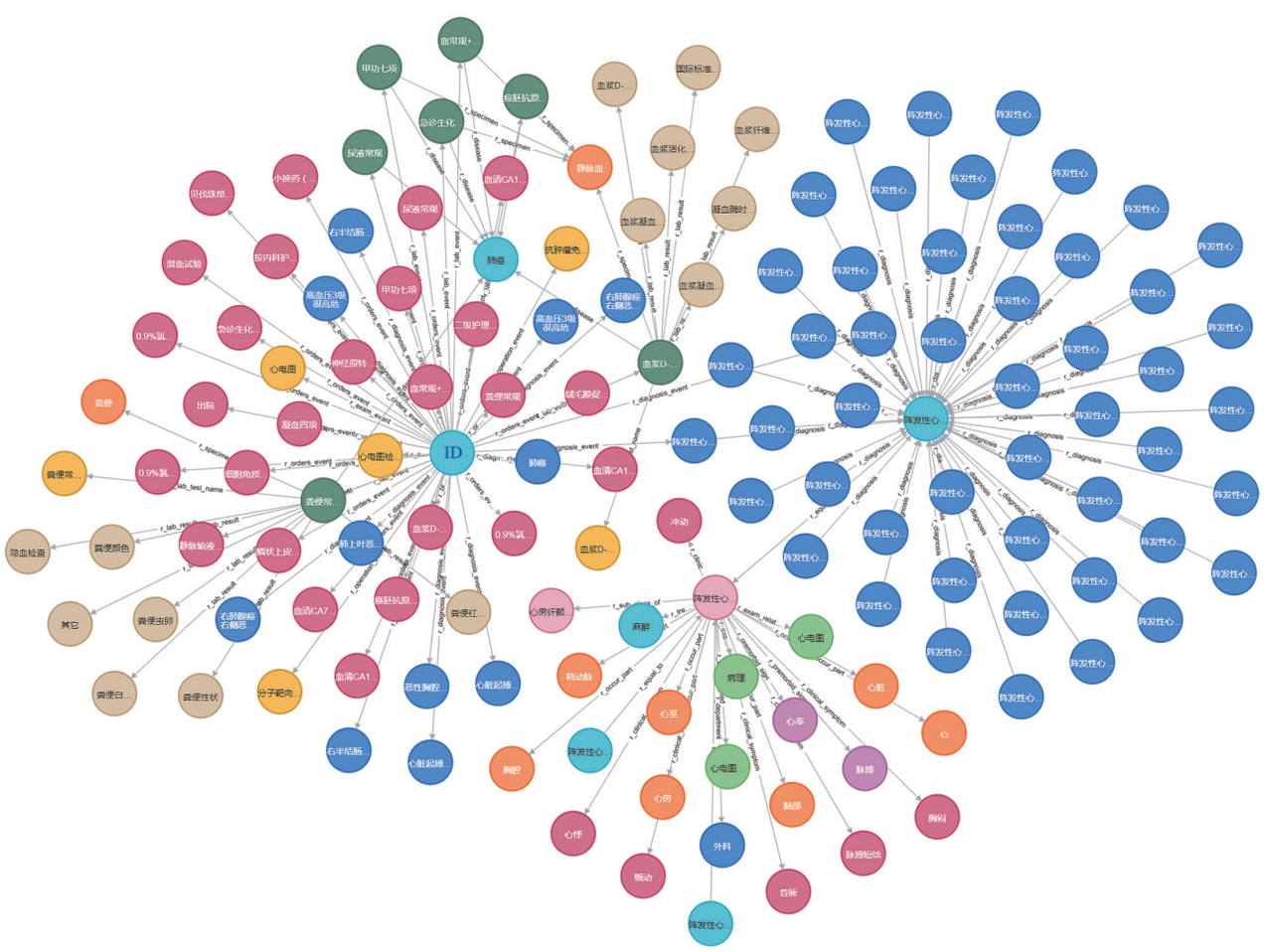

图5　涵盖医疗事件的知识图谱示例

Fig. 5　The examples of knowledge graph covering medical events

图 6　科研平台智能检索界面

Fig. 6　Intelligent search interface for scientific research platforms

大数据平台建设，各具特色，取得了一定应用成效。多数机构的数据治理过程主要由厂商完成，若涉及个性化需求时，需在厂商支持下才能完成。而本平台由解放军总医院主导研发，采用了模块化、组件化的设计理念，通过构建相对独立的数据处理工具，提升了平台的可扩展性和灵活性，既满足了临床数据检索与分析应用的需求，又依托平台提供的自然语言处理、变量中心等工具，支撑了数据工程师进行变量灵活定义和数据二次加工，有效降低了数据处理的复杂度和维护成本。

医疗数据具有复杂性高、数据量大、处理的实时性和精准性要求高等特点，传统数据处理技术只能适应结构复杂的小规模数据或结构简单的大规模数据，医疗数据的极大规模和充分集成将会是今后开展创新型医疗服务模式的基础，因此亟须汇聚大量人力、财力、物力，进行协同创新攻关和关键技术突破[18-19]。未来，本中心将继续探索区块链、联邦学习等关键技术，搭建数据安全条件下的跨区域、多中心数据应用协作模式，实现数据本地化联合分析，提高应用效能。

作者贡献　吴欢：论文撰写；车贺宾、陈媛媛：平台设计；乌日力格、王万玲：平台设计，论文校对和修订；何昆仑：论文指导。

利益冲突　所有作者声明无利益冲突。

数据共享声明　本论文相关数据可依据合理理由从作者处获取，Email：wh2531@126.com。

参考文献

1　郭强，王丛，衡反修. 医疗大数据平台建设机遇、挑战及其发展［J］. 医学信息学杂志，2021，42（1）：2-8.
2　师庆科，李楠，王觅也，等. 医疗健康大数据平台建设实施路径探索［J］. 中国数字医学，2023，18（1）：18-22.
3　杨亦含，胡元会，陈婷，等. 基于真实世界的中医临床科研大数据平台研发现状、不足及展望［J］. 中国中医基础医学杂志，2022，28（11）：1882-1886.
4　阮彤，邱加辉，张知行，等. 医疗数据治理：构建高质量医疗大数据智能分析数据基础［J］. 大数据，2019，5（1）：12-24.
5　潘军华，徐贝贝，李晓峰. 基于国家临床医学研究中心的临床研究大数据平台建设思考［J］. 北京医学，2021，43（9）：930-932.
6　席韩旭，张晨，张欣，等. 基于临床大数据的科研平台建设与应用探讨［J］. 医院管理论坛，2020，37（9）：67-68.
7　萧锘，叶琪，刘传丰，等. 区域临床科研大数据平台设计与实现［J］. 中国卫生信息管理杂志，2022，19（5）：673-680.
8　宋雪，王觅也，郑涛，等. 医院大数据平台建设难点及关键技术研究［J］. 中国卫生信息管理杂志，2024，21（2）：286-290.
9　徐骁，胡外光，陈敏莲. 临床科研一体化服务平台建设与应用探讨［J］. 医院管理论坛，2022，39（10）：68-73.
10　吕旭东，田琪，蔡海领，等. 临床科研数据库平台关键技术研究与实现［J］. 中国数字医学，2021，16（1）：23-29.
11　张弘政，刘迷迷，李琳，等. 基于通用数据模型的健康医疗大数据平台数据治理研究［J］. 医学信息学杂志，2022，43（6）：2-7.
12　刘辉，蔡宏伟，高娟娟，等. 大型综合性医院生物样本信息资源大数据科研平台的建设与应用［J］. 医学信息学杂志，2024，45（1）：77-82.

（下转第133页）