

临床专病数据库建设现状与思考^{*}

王耀国^{1,2} 李 鹏¹ 刘迷迷³ 龙思哲⁴ 李志健⁵ 周 毅³

(¹ 新疆维吾尔自治区人民医院 乌鲁木齐 830001 ² 新疆医科大学公共卫生学院 乌鲁木齐 830054
³ 中山大学中山医学院 广州 510080 ⁴ 中山大学附属第一医院信息数据中心 广州 510080
⁵ 长春理工大学 长春 130022)

〔摘要〕 目的/意义 为我国专病数据库建设、临床科研和医院信息互联互通等提供新思路。方法/过程 结合国内外医疗体系和医疗信息化发展历程阐述专病数据库发展现状，总结国内两种主流专病数据库构建模式，提出基于 SQL 集成技术和通用数据模型的专病数据库构建方案。结果/结论 该方案可快速构建专病数据库，并在数据层面实现互联互通。

〔关键词〕 专病数据库；临床科研；通用数据模型；互联互通

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.03.011

Current Situation and Considerations on the Construction of the Clinical Specialized Disease Database

WANG Yaoguo^{1,2}, LI Peng¹, LIU Mimi³, LONG Sizhe⁴, LI Zhijian⁵, ZHOU Yi³

¹ People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi 830001, China; ² School of Public Health, Xinjiang Medical University, Urumqi 830054, China; ³ Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China; ⁴ Center of Information Technology and Statistics, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China; ⁵ Changchun University of Science and Technology, Changchun 130022, China

〔Abstract〕 **Purpose/Significance** To provide new ideas for the construction of specialized disease database, clinical scientific research and hospital information interconnection in China. **Method/Process** Combined with the development history of domestic and foreign medical systems and medical informatization, the paper expounds the development status of specialized disease database, summarizes the two main construction modes of specialized disease database in China, puts forward the construction scheme of specialized disease database based on structured query language (SQL) integration technology and common data models. **Results/Conclusion** The solution can quickly build a specialized disease database and realize interconnection at the data level.

〔Keywords〕 specialized disease database; clinical scientific research; common data model; interconnection

〔修回日期〕 2023-08-15

〔作者简介〕 王耀国，硕士，发表论文 2 篇；通信作者：周毅，博士，教授，博士生导师。

〔基金项目〕 国家重点研发计划（项目编号：2022YFC3601600）；广东省广州市科技计划项目（项目编号：202206010028）；新疆维吾尔自治区人民医院院内课题（项目编号：20230117）。

1 引言

专病数据库发展与国家政策和临床医学需求密切相关。国内外不同的医疗体系、医疗信息化历程对专病数据库建设和发展方向产生不同影响。2023 年 3 月 7 日国务院宣布组建国家数据局,从国家层面协调推进数据基础制度建设。目前,我国卫生健康信息化领域顶层设计体系基本形成^[1],为健康医疗大数据互联互通指明发展方向。然而,各地医疗信息化发展不均衡且底层数据缺少统一的行业标准,数据互操作性较差,原始医疗健康数据需要标准化处理后才可实现互联互通。随着医学科学的发展,精准医学和临床亚专业逐渐兴起,疾病研究越来越深入,临床研究更加依赖某一类甚至某一种疾病的病例数据。由于医疗信息系统建设时期及要求的差异,不同信息系统底层数据标准不一致,数据互通和整合存在问题,虽然存有大量临床数据,却无法直接支持临床研究。提高临床科研数据可用性是专病数据库产生的直接原因,采用专病数据库提高数据的互操作性也是实现数据互联互通的方式之一。目前国内临床各分支学科对专病大数据建设需求较强,希望通过建立专病数据库对目标患者群体内在生物医学信息及电子病历数据进行整合与分析利用,并发掘和提取有价值的信息,以提高专科疾病医疗服务质量及治疗效果,进而规避和降低医疗风险,减少医疗成本等^[2-4]。

2 专病数据库发展现状

2.1 国外专病数据库发展现状

发达国家健康医疗数据信息开放共享程度较高,医疗数据库建设较规范,目前已经形成一套比较完善的以大数据管理为核心的医疗数据管理体系。以“disease and database”为关键词,在 PubMed 中搜集 2020—2023 年外文文献共 6 633 篇。其中大多数针对特定疾病的国际临床研究建立了大型队列研究,为特定研究目的收集数据,从医疗服务和调查中收集特定疾病数据。这些数据大多是具

有前瞻性、经过精心过滤和处理的,而且数据数量相对较少。欧美等国家已经建立国家级专病数据库^[5-6],Cohen J A 等^[7]和 Okada M^[8]从医疗器械、生物制药公司抽取相关医疗数据,构建大规模数据库;Tsuneishi Y 等^[9]和 Kyoung D S 等^[10]利用日本国家健康保险索赔和特定健康检查数据库、韩国国民健康保险服务、健康保险审查与评估数据库进行数据分析。

美国约在 2016 年完成医疗信息大规模互联互通工作,为构建国家级专病数据库提供平台基础。日本医疗模式与美国有较大不同,其老龄化率全球最高,但卫生支出占国内生产总值比重远低于美国。日本医疗主体以私人为主,国内实行严格的分级诊疗制度,全民医保,公共财政负担大部分医疗费用,这在一定程度上导致日本临床医疗数据与保险公司数据关联密切。中国与国外医疗信息模式本质不同,但路径有相似之处,均是由基础运营管理信息化向临床业务逐渐渗透,并在新兴技术加持下,以提升用户与患者体验为核心,构建智慧医疗生态圈,实现医疗服务模式革新^[11]。

2.2 国内专病数据库发展现状

以“专病数据库”为关键词,在中国知网、万方数据等网站检索收集相关文献共 52 篇。总结发现我国专病数据库发展程度和方向受到国内医疗体系和医疗信息化程度的共同影响,目前处于快速发展阶段。其建设以三甲医院为主,大多数正处在以数据为导向,建立院内科研数据平台及区域互联互通医疗平台阶段。目前我国国家级重大疾病专病平台数量较少,缺少业内公认、标准的设计和开发方法^[12]。该领域工作发展状况不均衡,上海、广州、北京等地医院专病数据库建设起步较早^[13],如上海交通大学转化医学研究院院长陈赛娟院士于 2022 年提交的《关于发展重大疾病专病数据库》提案;中山大学孙逸仙纪念医院院长宋尔卫院士牵头的“中国乳腺癌标准数据库”建设,已取得实质性进展;南方医院侯凡凡院士围绕肾脏疾病防治发展需求,加强肾脏病专病大数据标准化建设等。但西北等地医院专病数据库发展程度和规模相对落后。

3 专病数据库构建流程与主流构建方案

3.1 专病数据库构建流程

专病数据库通用构建流程大致分为 3 部分：数据整合、数据处理和数据安全。专病数据库中多源异构数据来源包括医院信息系统、电子病历系统、检验系统、影像系统等。数据整合一般采用抽取、转换和加载（extract – transform – load，ETL）技术。数据处理一般采用自然语言处理技术、机器学习算法和深度学习模型，然后通过可视化页面进行展示。数据安全则贯穿始终，包括确保数据不泄漏、数据脱敏，采用数据可用不可见、可见不可取等原则，见图 1。

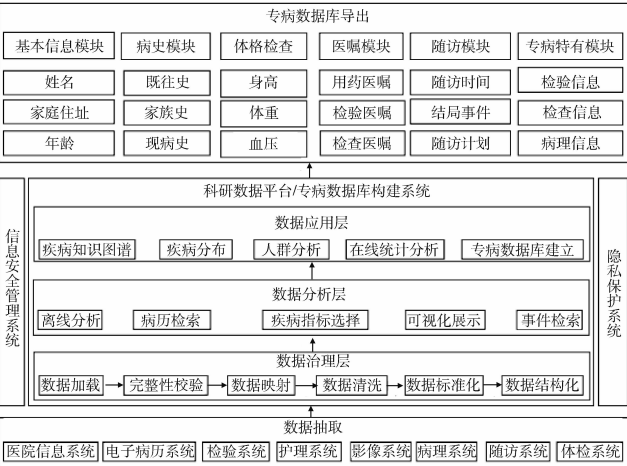


图 1 专病数据库通用体系架构

专病数据库抽取方案主要是根据患者主索引，利用视图、集成平台等方式实时或定时抽取数据。根据数据处理难度，依次大致分为 3 类：第 1 类为结构化数据，如数值型数据，可以直接存储，如人口学、检验等信息；第 2 类为形式结构化但内容非结构化数据，如“一诉五史”等长文本，要从中提取特定的医疗事件，一般先用自然语言技术再进行算数运算。以病程提取为例，需要从一段疾病史描述中提取疾病开始时间，再结合出院时间确定病程，之后再填充到指定字段^[14]；第 3 类为非结构化数据，如脑电信号数据、图像数据等。专病数据库一般采用关系数据库存储，底层数据存储的逻辑结

构为分表存储，同时通过患者主索引（身份证、各种检查号等）实现各模块间每个患者相关信息的关联^[15-16]。专病数据库大致分为两大类模块：通用信息模块，包括患者人口学信息模块、病史模块（既往史、现病史、家族史、主诉、病程、用药史）、体检模块等；专病数据信息模块，包括检验模块、检查模块、病理模块等。

3.2 国内专病数据库主流构建模式

第 1 种基于业务系统构建专病数据库。先根据临床科室确定目标数据集，再直接从医院内各种业务系统中抽取数据^[17]，主要用于构建小数据量和单病种专病数据库。该方案搜集的临床数据主要是数值和文本，对检查影像等临床数据的处理一般仅搜集其文本描述信息，如检查时间、检查所见、检查结论等，不整合影像结果。主要有两方面应用：一是为临床科研提供数据支持；二是根据专病数据库数据质量报告进行病历质控。

第 2 种基于科研数据平台构建专病数据库。该方案是目前医院建设专病数据库的主流方案。首先，建立面向全院数据的科研数据平台，通过数据库备份或者数据同步技术持续抽取全院数据。其次，利用自然语言处理技术完成数据清洗、数据映射、数据存储、可视化等，利用科研数据平台可以实现对医院所有接入系统的数据整合、数据管理、统计分析、可视化展示、智能检索、队列发现及人工智能辅助决策等。最后，在科研数据平台的基础上抽取某种疾病的全部临床数据。

4 问题讨论与方案探索

4.1 问题讨论

医院信息系统数据库与专病数据库最大的区别在于前者面向临床业务，作为联机事务处理的组成部分；而后者面向临床科研，重点在于多源异构数据整合，对临床数据的获取、清洗允许存在一定时延。目前我国医疗信息化领域的痛点在于海量临床数据闲置，无法有效地支撑医院信息化发展和临床科研。其主要原因是医院内部各种信息系统存在

“数据孤岛”问题，各医疗数据库之间的数据互操作性较差。专病数据库就是为了解决多源异构数据的交互问题。目前国内构建专病数据库的方案已商业化落地，主要聚焦于临床数据的可用性，为临床科研提供有效数据支持。但本质上仍是一套独立的信息系统，处理后的数据与医院信息系统数据交互效率较低，有的甚至不交互，并未在真正意义上解决问题。

4.2 方案探索

4.2.1 总体架构 本研究基于上述专病数据库构建模式和流程拟提出一种基于结构化查询语言（structured query language, SQL）集成技术和通用数据模型（common data model, CMD）的专病数据库构建方案，旨在借助医院现有平台探索在数据层面实现互联互通。由专病数据库的构建流程可知，其构建过程存在 3 大关键点和难点：一是从各种医院信息系统中抽取数据；二是多源异构数据的标准化存储；三是数据安全的保障。因此，该方案首先收集医院信息系统的底层数据存储架构；其次构建可动态配置 SQL 等查询语句的数据检索平台进行数据抽取；再次基于通用数据模型通过编程手段实现数据抽取过程中的标准化映射，提高数据的互操作性，形成初步专病数据库；最后结合临床研究目的有针对性地对抽取数据进行清洗、编码、统计分析和临床建模，见图 2。

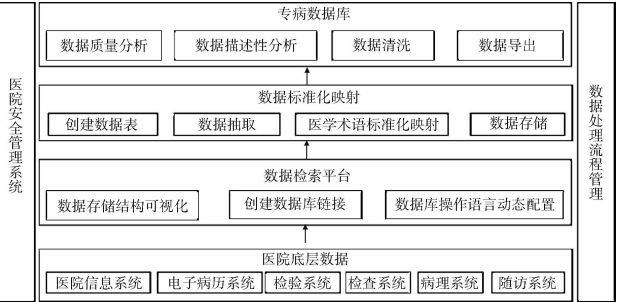


图 2 基于 SQL 集成和通用数据模型构建专病数据库

4.2.2 基于 SQL 集成技术的数据抽取模型 目前国内医院主流数据存储方式为关系数据库存储，如 Oracle、MySQL 等，为基于 SQL 集成技术的数据抽

取提供平台基础。不同疾病的数据抽取条件大同小异，字段数量可能不同但是来源于同一系统。SQL 抽取语法高度类似，因此通过 SQL 语句集成并根据不同疾病进行微调，即可快速构建一个专病数据库。

4.2.3 基于医学术语映射平台的数据标准化存储 医院信息互联互通客观要求医学术语标准体系之间融合、映射和翻译对照，但不同医学术语系统之间的映射工作量巨大，因此业界提出一种集成各种医学术语系统的医学术语平台概念。通用数据模型具有统一的医学概念表达形式，包含标准化词汇表、标准化元数据、标准化临床数据表、标准化健康系统数据表、标准化健康经济表和标准化派生元素 6 类，共 39 张表。通用数据模型是数据标准化的核心，目的是将不同数据库包含的数据转换为通用格式以及应用统一术语^[18-25]。基于通用数据模型适配多源异构数据，使用通用医学信息标准对原始临床数据进行标准化，可以减少后期人工对照时间，提高数据完整性和一致性以保证研究数据质量。

该方案与既往主流专病数据库构建方案的不同之处在于依托现有信息平台，不需要额外的硬件平台；此外不同于科研数据平台对所有临床数据的整体抽取，可根据医院信息系统的底层数据存储架构创建基于 SQL 集成、可动态配置查询语句的数据检索平台，抽取指定数据，对医院而言存储负担较小。同时利用医学术语平台等工具提升数据的互操作性，实现数据层面的互联互通，避免硬件平台限制，为医院信息部门自行处理临床数据、深度参与临床科研提供一种思路，为智慧医院和区域医疗平台建设提供数据基础。该方案的缺点是对医院信息化人才队伍要求较高，虽然实现了构建过程重用，但是构建专病库时需要根据临床业务和底层数据逻辑进行微调，因此技术学习成本高，需要既懂临床业务又熟悉信息化的复合人才。

5 结语

随着我国医疗信息化的发展，医院信息互联互通

将是大势所趋。本研究提出一种专病数据库构建方案, 在提高临床数据可用性的同时, 实现数据层面的互联互通并提高临床数据的互操作性。虽然该方案在技术上可行且已在医院完成多个专病库的构建, 但其通用性和实用性有待在临床实践中进一步验证。高质量的构建和应用仅依靠医院某几个部门已不能满足实际需求^[26]。为了提升质量, 专病数据库的构建应由医院多部门、多学科人员共同参与^[27]。医院信息部门与临床科研人员在数据处理、分析、建模等方面的深度合作也可以提高科研效率。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 董方杰, 胡建平, 吴士勇. 我国卫生健康信息互联互通 2.0 技术特征研究 [J]. 中国卫生信息管理杂志, 2023, 20 (1): 1-6.
- 薛万国, 乔岫, 车贺宾, 等. 临床科研数据库系统的现状与未来 [J]. 中国数字医学, 2021, 16 (1): 2-6.
- 刘迷迷, 杜国霞, 周毅, 等. 专病数据库建设与应用研究 [J]. 医学信息学杂志, 2021, 42 (11): 81-86, 93.
- 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): 1-8.
- SAYERS E W, BECK J, BOLTON E E, et al. Database resources of the national center for biotechnology information [J]. Nucleic acids research, 2021, 49 (D1): D10.
- ROGER V L. Epidemiology of heart failure: a contemporary perspective [J]. Circulation research, 2021, 128 (10): 1421-1434.
- COHEN J A, TROJANO M, MOWRY E M, et al. Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment [J]. Multiple sclerosis journal, 2020, 26 (1): 23-37.
- OKADA M. Big data and real-world data-based medicine in the management of hypertension [J]. Hypertension research, 2021, 44 (2): 147-153.
- TAKEKUMA Y, IMAI S, SUGAWARA M. Clinical research using the large health insurance claims database [J]. Journal of the pharmaceutical society of Japan, 2022, 142 (4): 331-336.
- KYOUNG D S, KIM H S. Understanding and utilizing claim data from the Korean National Health Insurance Service (NHIS) and Health Insurance Review & Assessment (HIRA) database for research [J]. Journal of lipid and athero-

sclerosis, 2022, 11 (2): 103.

- 张雪峰. 我国医疗信息化行业现状及商业模式分析 [J]. 投资与合作, 2022 (7): 31-33.
- 陈赛娟. 发展重大疾病专病数据库意义重大 [J]. 科技创新与品牌, 2022 (3): 20.
- 袁骏毅, 潘常青, 戴锦杰, 等. 基于数据治理的先心病专病数据库建设研究 [J]. 中国卫生质量管理, 2022 (2): 77-80.
- 吴燕秋, 黄伟, 刘慧鑫, 等. 医院创伤专病数据库建设与实践 [J]. 医院管理论坛, 2021, 38 (5): 79-82.
- 孙颖, 李超峰, 林丽, 等. 鼻咽癌专病科研数据库建设与应用 [J]. 中国数字医学, 2021, 16 (1): 7-12.
- 屈媛圆, 高琳娜, 宋欢. 结直肠癌专病数据库的发展与展望 [J]. 中国普外基础与临床杂志, 2021, 28 (12): 1543-1545.
- 陈嘉旖, 陶蓉. 基于临床大数据的肺癌专病数据库建设 [J]. 中国数字医学, 2021, 16 (11): 39-42.
- 赵前前. 基于大数据科研平台的专病数据库建设及应用 [J]. 中国数字医学, 2020, 15 (12): 89-92.
- 李慧杰, 张晴晴, 刘瑞红, 等. 大数据背景下临床专病数据库建设实践与思考 [J]. 中国卫生事业管理, 2020, 37 (8): 574-576, 591.
- 范小红, 袁骏毅, 汪刚. 基于 AHP 和 TOPSIS 法的肺部肿瘤数据库优化设计的研究 [J]. 中国数字医学, 2019, 14 (4): 88-90, 81.
- 荣雯雯, 汪刚, 朱其立. 基于人工智能的病历后结构化专病数据库在临床研究中的价值探讨 [J]. 上海交通大学学报 (医学版), 2020, 40 (7): 995-1000.
- 翟孝庭, 黄东宗, 李鸿波, 等. 颞下颌关节紊乱病专病数据库的设计与初步构建 [J]. 口腔颌面修复学杂志, 2021, 22 (5): 354-360.
- 惠美蓉, 仇鹏, 王世鹏, 等. 研究型下肢动脉硬化闭塞症专病数据库的建立及应用 [J]. 中国血管外科杂志 (电子版), 2020 (3): 248-251.
- 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, 13.
- 侯丽, 洪娜, 李露琪, 等. OHDSI 通用数据模型及医学术语标准国内应用现状分析 [J]. 医学信息学杂志, 2020, 41 (2): 2-10.
- 姬卫东, 李琳, 张振, 等. 互联互通背景下医疗数据治理面临的问题与对策 [J]. 中国数字医学, 2021, 16 (11): 6-10.
- 梁文全, 郝洪庆, 乔岫, 等. 基于单中心万例胃癌专病数据库的胃癌外科临床病理信息分析 [J]. 中华胃肠外科杂志, 2020, 23 (11): 1051-1058.