• 医学信息技术 •

基于数据治理的脑血管专病数据库建设实践*

连万民 段文舟 张冬平 王博涵

（广东省第二人民医院 广州 510317） （广州知汇云科技有限公司 广州 510000）

〔摘要〕 基于广东省第二人民医院脑血管专病数据库建设实践，介绍基于数据治理的专病数据库系统架构，探讨专病模型、后结构化处理、多维数据管理、智能搜索等关键技术应用，并阐述专病科研平台建设成果。

〔关键词〕 数据治理；专病数据库；脑血管病

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.05.011

Construction Practice of the Cerebrovascular Disease Database Based on Data Governance *LIAN Wanmin, DUAN Wenzhou, Guangdong Second Provincial General Hospital, Guangzhou 510317, China; ZHANG Dongping, WANG Bohan, Guangzhou AID Cloud Technology Co. Ltd., Guangzhou 510000, China*

〔**Abstract**〕 Based on the construction practice of the cerebrovascular disease database in the Guangdong Second Provincial General Hospital, the paper introduces the system architecture of the cerebrovascular disease database based on data governance, discusses the application of key technologies such as specialized disease model, post－structured processing, multidimensional data management, intelligent search, and expounds the achievements of the construction of the specialized disease research platform.

〔**Keywords**〕 data governance; specialized disease database; cerebrovascular disease

1 引言

近年来以电子病历为核心的医疗机构信息化建设得到大力推动，“互联网 + 医疗健康”应用日渐广泛，医院诊疗数据、检查检验数据、健康人群体检数据、队列随访数据、药物使用数据、病理和影像数据、基因组学等健康大数据快速增长，推动医疗健康领域进入“大数据”时代[1]。基于真实世界的大数据研究分析成为当下研究热点。然而，不少医院院内虽然具有良好的信息化基础，但是数据质量不高、数据开发难度较大，缺乏统一数据开发平台[2]，导致医务人员在推动疾病诊断、治疗、预后的研究和发展方面缺乏相应数据及技术支持，医院积累的宝贵经验无法得到高效分享，医疗证据不能得到合理应用。

脑血管病是全球性公共卫生问题。2019 年全球疾病负担课题组（Global Burden of Disease 2019, GBD 2019）数据显示，脑血管病是 204 个国家和地

〔修回日期〕 2022－10－11

〔作者简介〕 连万民，高级工程师，发表论文 9 篇。

〔基金项目〕 国家重点研发计划项目“养老助残友好智慧健康宜居环境体系构建与应用示范”（项目编号：2021YFC2009400）。

区居民死亡和过早死亡的主要原因之一[3]。根据国家统计局数据，近年来脑血管病死亡人数一直处于高位，2017—2021 年城市、农村脑血管病死亡人数占总死亡人数的平均比重高达 20.86% 和 23.92%[4]，且有逐年升高趋势；与此同时，随着我国人口老龄化不断加剧，到 2022 年我国 65 岁以上人口比例已达 14.86%，这一疾病负担将日趋严重[5]。脑血管病以其发病率高、复发率高和致残率高的特点成为严重阻碍我国社会经济发展的重大疾病[6]。脑血管病的治疗需要一体化、全链条干预，一般急性期在神经内科治疗，病情平稳后即可进入康复理疗科治疗。但是，医院信息系统常存在数据资源无法共享以及多系统、多业务存储底层数据结构不统一等问题[7]，导致神经内科与其他相关科室之间的临床数据处于信息孤岛状态，无法充分挖掘其价值。

发达国家围绕临床与科研已经广泛开展脑血管病相关数据库建设与应用研究。例如，欧洲建立卒中数据库对急性脑血管病患者的人口统计学特征、危险因素、卒中严重程度和治疗效果进行研究[8]，美国建立心房纤颤导管消融术后 30 天急性脑血管意外发生率和预测因素研究数据库[9]，均取得较好效果。因此基于全院数据治理框架，建设脑血管专病科研数据库，利用数据挖掘与人工智能等技术掌握不同病因脑血管病发生、发展和转归特点，最终阐明疾病发病机制进而辅助临床决策，对延缓病程进展、准确预测并发症发生及死亡风险、早期治疗、干预等具有临床和社会意义[10]，同时对提高人均预期寿命、降低重大慢性病过早死亡率具有重要现实意义。

2 建设实践

2.1 系统架构

2.1.1 整体架构 提高临床数据可及性和可用性是临床科研数据库平台需要解决的核心问题[11]。数据治理是提升数据质量和可利用性的重要手段。但是针对不同应用场景和数据基础，数据治理的总体框架和核心任务则反映各自专业需求特点，各不相同[12]。因此，广东省第二人民医院以建设高效、灵活、方便、安全、一体化的科研专病数据库系统平台为目标[13]，结合现有数据基础，借鉴治理基础层、数据加工层、价值体现层 3 层治理框架，进行临床科研数据治理，通过算法、逻辑、规则、功能模块，执行标准化、元数据与主数据管理、数据建模、数据采集、数据归集、数据加工、数据挖掘、数据展示、质量控制等核心任务[14]，见图 1。

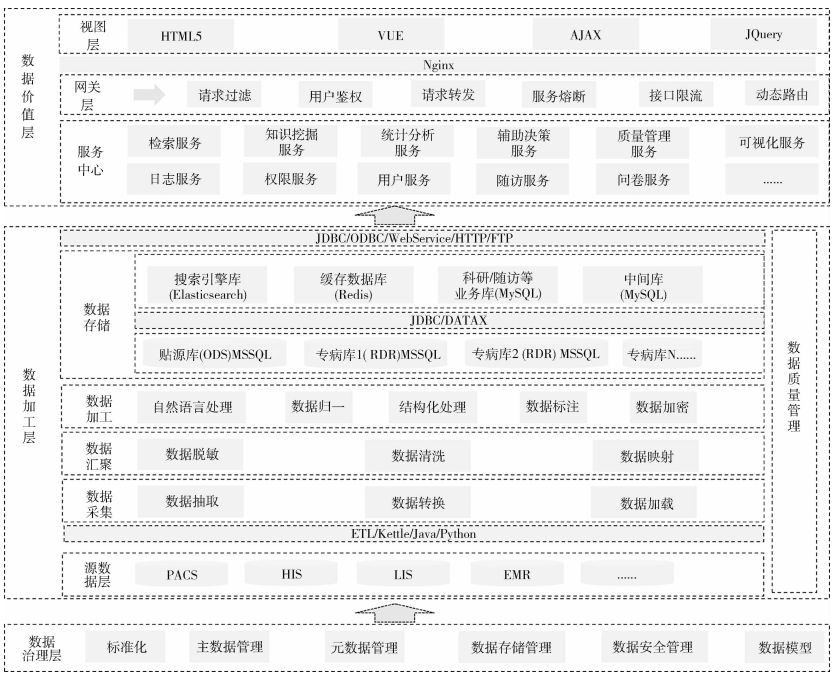

图 1 系统整体架构

2.1.2 基本工作流程 影像存储与传输系统（picture archiving and communication system，PACS）、医院信息系统（hospital information system，HIS）、电子病历系统（electronic medical record，EMR）、实验室信息管理系统（laboratory information management system，LIS）等业务系统和临床数据库（clinical data repository，CDR）按照电子病历数据集、国际疾病分类法（international classification of diseases，ICD）、观测指标标识符逻辑命名与编码系统（logical observation identifiers names and codes，LOING）、医院信息互联互通成熟度测评标准完成业务数据标准化，通过数据仓库技术（extract - transform - load，ETL），根据实时性要求分别从各业务系统或临床数据库中抽取数据，经清洗、转换、加载等初步加工处理形成原始病历库。然后通过数据映射、自然语言处理（natural language processing，NLP）、正则规则等深度加工后与结构化数据按照主题域构建科研病历库，再根据专病模型对数据进行逻辑归集形成各专病库。基于科研病历库，系统为用户提供知识挖掘、全文检索、复杂统计、质量管理等各种应用输出。

2.2 关键问题

2.2.1 基于标准化的专病模型 专病数据库本质上是对多个分散异构业务系统的诊疗数据通过 ETL 进行形式和内容上的二次加工，使其符合科研数据库的数据模型。不同专病数据库一般采用自定义的数据模型，在对外数据共享、建立多中心专病库时需要耗费大量资源进行对接改造和数据映射。为解决这些问题，需要引入数据标准，建立标准化的临床数据模型、医学术语、编码系统[15]。广东省第二人民医院脑血管专病库大部分数据来源于 CDR。其中数据在进入前已遵循电子病历数据集、ICD、LOINC、医院信息互联互通成熟度测评等标准规范在数据层面进行标准化。专病模型的确立与研究目的、建库工作量和后期扩展性高度相关。健康医疗数据科学与信息学组织（Observational Health Data Sciences and Informatics，OHDSI）提出的观察医疗结果合作项目通用数据模型（observational medical outcomes partnership common data model，OMOPCDM）是一个为医学数据标准化而设计的数据模型[16]。借鉴该模型，结合国情与项目目标对数据的需求，进行专病库数据模型设计，就既往科研病历报告表单（case report form，CRF）与研究课题所需数据项进行深度沟通，最终确认模型构成。再根据模型搭建专病数据库，对数据中心以及业务系统的数据进行抽取、清洗并加载至数据库中，对部分数据项实现标准化清洗，对关注的医嘱药品、检验、诊断信息等数据进行归一化处理，对多来源数据项进行关联和逻辑计算。根据专病库实时性要求，通过 ETL 工具实现数据自动增量，对增量流程进行监控，实现数据量统计、日志记录、报错智能提醒等功能。对数据溯源关系、数据处理脚本进行封装，保证 ETL 流程透明化。编写数据质量脚本进行专病数据量统计、完整度计算、多来源数据项一致性校验，实现专病库数据质量控制。

2.2.2 后结构化处理 为提高数据质量，专病数据库通常会在临床业务信息系统通过结构化模板等方式进行数据的前结构化，但是临床表达与使用习惯等不同会导致部分数据不能实现结构化。然而临床科研关注的数据往往包括非结构化数据，如转科检查和转科病历等医疗文书，因此需要通过 NLP 进行后结构化处理。考虑到传统 NLP 医生标注的工作量和成本，脑血管专病库使用超过 5 万篇电子病历数据，基于 TensorFlow 框架，主要采用无监督深度学习方法训练得到垂直领域专病语言模型。模型以改造后的 BERT 预训练语言表征模型为基础，结合相关指标信息（包括指标名称、同义词、数据来源等特征），自动抽取与指标相关的病历原文，然后结合内部医疗知识库，利用命名实体识别及关系抽取算法自动生成规则，从而完善语义规则引擎及知识库，最终完成指标的自动提取和后结构化，见图 2。

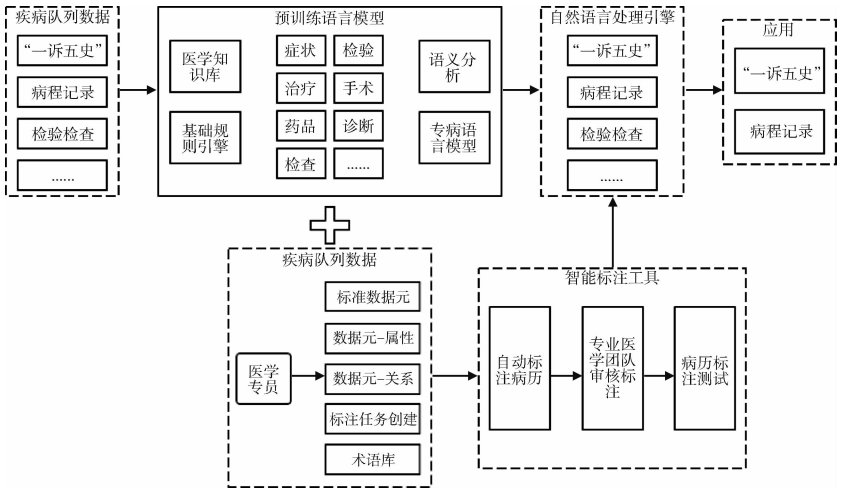

图 2 后结构化工作流程

利用该病历后结构化流程，所有枚举型指标都将跳过人工标注，直接通过预训练模型自动后结构化。模型在自动学习过程中不断完善知识库，持续提高效率和精度。此外，结合临床医学特点进行语义化分词，将分词后的结构以临床医生熟悉的专科词汇进行存储，便于在科研、临床辅助过程中快速获取关键病历信息，如症状、特征值、阳性特征等，最大化、最快速地为科研提供临床参考资料。

2.2.3 *多维数据关联* 最大程度地从原始医疗数据中自动关联和提取病历数据是减轻临床科研人员数据整理工作量的关键[17]。由于历史原因，同一患者在医院可能有多个身份标识，同时医疗数据包括文本、图片、视频、表格等多种数据类型，具有多维特征，在数据抽取时，容易遗漏，造成数据不完整。因此，通过建立患者主索引（enterprise master patient index，EMPI），应用特定算法将不同业务系统所提供的患者标识信息重新组织，生成同一患者的唯一标识编码企业级患者主索引识别码（enterprise master patient index_identity，EMPI_ID），根据此编码能找到分布在各业务系统中的患者所有医疗信息，同时消除重复的患者数据，实现跨系统信息检索与共享[18]。EMPI 同时提供患者信息检索服务，提供给其他应用程序访问患者基本信息；考虑到对异构平台的支持，消除系统平台的环境差异性等因素，EMPI 通过接口对外提供服务，例如医院随访系统可以传入患者关键信息（姓名、性别、出生日期、身份证号、联系电话等），通过调用 EMPI 服务接口返回或生成对应的 EMPI_ID，各业务系统都可以通过 EMPI 提供的接口来检索相关患者用户信息[19]。专病库在 ETL 中使用患者主索引服务，通过患者姓名和身份证进行精确匹配，通过姓名、性别、出生日期、联系方式等属性的权重进行模糊匹配，合并患者标识，生成或获得 EMPI_ID，然后以就诊时间串联患者历次就诊记录形成纵向时间轴，横向以每次就诊的流水号关联各类型就诊数据，最终实现患者多维数据关联，并可通过时间轴上的链接调阅病历、影像等数据，见图 3。

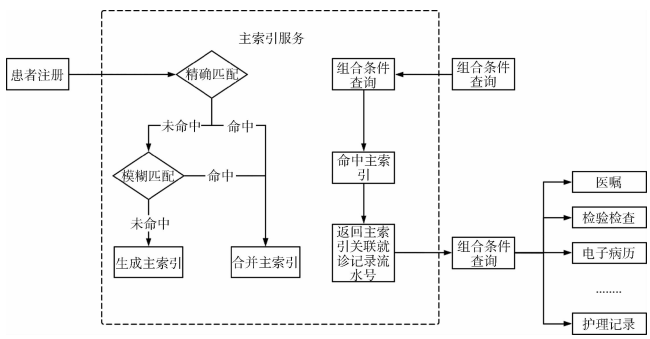

图 3 主索引服务

2.2.4 *智能检索* 对已确定的科研主题，通过特定条件精准筛选病历数据，建立科研队列，对队列数据进行分析，按临床需求设计和生成 CRF 表单并导出数据，提供数据分析工具，采用统计分析方法对数据的分布状态、数字特征和随机变量之间的关系进行

定性或定量估计和描述[20]。医生也可通过指定检索条件快速归集病历，形成临时队列，通过对队列的分析总结产生新的科研课题。因此专病库需要强大的搜索引擎，能高效实现专病库各数据项多重组合条件的检索和病历文书的全文检索。考虑到脑血管专病库数据量和对检索效率的要求，采用广泛应用的 Elasticsearch 数据分析引擎，通过对底层开源库 Apache Lucene 的封装，实现对每个数据项的索引和搜索。首先从科研数据库中抽取数据写入搜索引擎 Elasticsearch Index；系统建设初期使用全量抽取，之后通过增量方式进行抽数。完成索引后，当系统接收到用户检索条件的请求，自动匹配定义的数据元，并利用系统自身逻辑程序封装成 Elasticsearch 的 DSL 语句；而后基于 Elasticsearch 的底层能力，根据 DSL 语句从 Elasticsearch Index 中检索。如检索到数据，进行相关处理（如脱敏等操作）后，返回给用户。用户可以预览数据和下载 CRF，见图 4。相关数据结果可在智能统计平台进行分析。

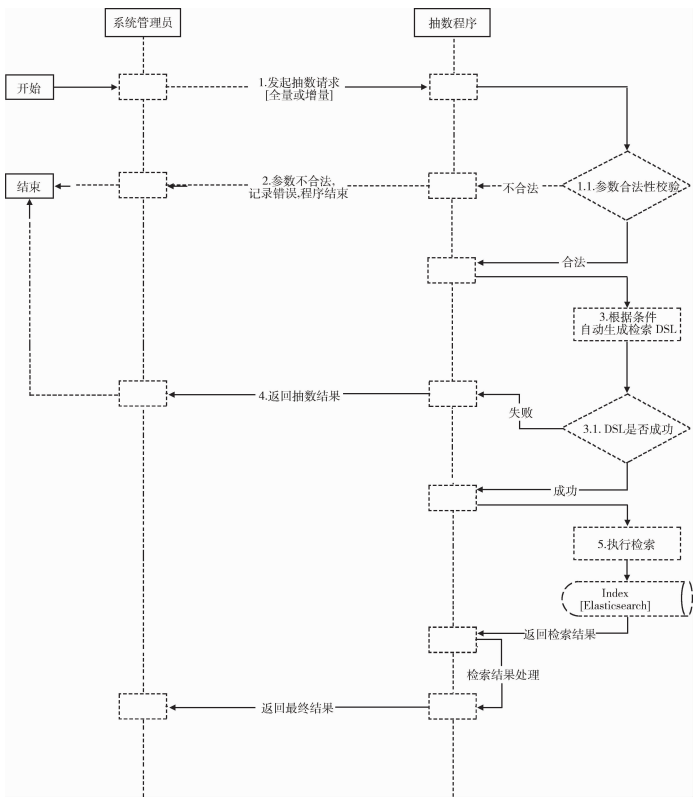

图 4 数据检索流程

3 建设成果

该脑血管专病数据库于 2019 年 12 月开始建设，2020 年 6 月上线，其中纳入近 8 年以脑血管病为主诊断的患者 38 391 例。通过数据映射，应用同义词归一等数据治理手段，将住院及门诊不规范诊断 24 706 种归一映射出主诊断为脑血管病、脑梗死、高脂血症、脑梗死后遗症等 241 种诊断。

专病库采集数据元按业务域分为 24 类，共计 1 188 项，见表 1。其中包含结构化数据 747 项，361 项通过映射实现值域与数据字典的一致性；非结构化数据 441 项。

表 1 数据概览

序号	业务域	获取方式	数据元数量	记录数
1	人口学信息	直接映射	29	943 000
2	就诊信息	直接映射	35	3 112 986
3	诊断信息	直接映射	76	7 562 518
4	一般检验	直接映射	55	76 263 337
5	微生物检验	直接映射	14	1 357 838
6	检查报告	直接映射	82	1 300 948
7	医嘱记录	直接映射	210	46 773 526
8	门急诊病历	直接映射	12	239 612
9	住院病历	直接映射	8	2 661 983
10	入院记录	后结构化	302	2 166 458
		直接映射	20	290 850
11	手术记录	后结构化	38	9 122
		直接映射	38	168 388
12	生命体征	直接映射	8	10 963 656
13	出院小结	后结构化	8	412
		直接映射	12	364 695
14	绿道信息	直接映射	29	11 024
15	费用信息	直接映射	13	23 353
16	脑血管造影	后结构化	17	98
		直接映射	26	49 189
17	日常病程记录	后结构化	17	4 351
		直接映射	8	703 660
18	首次病程记录	后结构化	11	43
		直接映射	12	332 128
19	术后首次病程记录	后结构化	14	127
		直接映射	9	16 536
20	术前情况表	直接映射	14	39 710
21	术中记录信息	直接映射	21	1 467 714
22	颅脑磁共振检查	后结构化	10	17 078
23	超声检查	后结构化	24	41 708
24	观察项目	直接映射	16	3 585

其中，对脑血管病诊疗核心的入院记录、手术记录、脑血管造影、颅脑磁共振等，需要后结构化指标 441 个，临床专家团队对每类 100 例报告进行标注，技术团队经过 1 万份样本训练后，完成以上指标的自动化采集、清洗、治理及可视化。为脑血管病等复杂疾病诊疗数字化提供重要参考依据。

此外，根据脑血管病诊断特点并结合临床研究应用方便快速检索数据的需求，令“诊断条件”按“前后循环”“血管定位”“解剖定位”“定性诊断”平铺陈列，“影像信息”按“检查类型”“解剖部位”“血管部位”“病变性质”“灌注成像”平铺陈列，方便临床科研工作人员快速定位相关患者队列，还可以通过检索频次、常用检索匹配逻辑固定成检索项，快速锁定队列。目前，已建立脑梗－丘脑、脑梗－后循环等多个研究队列，通过指定条件检索入组病历，为科研提供数据支持，也可以通过对队列数据的分析研究挖掘新的科研课题。

4 结语

以数据治理的理论框架为指导，通过临床调研、专病数据模型建立、后结构化、数据抽取、智能检索等技术实践，建立脑血管专病数据库，为脑血管专病科研队列管理、临床回顾性研究、数据建模和相关性分析提供有力数据支撑。

在数据库建设过程中遇到一些问题需要结合具体情况制定相应解决方案。例如卒中绿色通道患者缺失较多院前急救信息，通过对院前急救系统的改造，使用平板快速录入关键信息，使用患者主索引关联就诊记录，直接从院前急救系统提取数据项。再如部分后结构化数据项经过多次标注和算法优化效果仍然不好，通过对病历模板复杂度和医生书写习惯的分析，使用前结构化方式对部分模板进行改造，同时进一步优化 NLP 算法，在平衡医生病历书写工作量、满足科研需求的前提下提高数据准确性。在后续数据分析建模过程中，存在部分数据项为文字描述型无法进行分析，以及连续型变量中掺杂文字符号无法进行量化等问题，增加异常值处理模块，自动分析数据项类型，将描述型变量按关键词转换为多分类变量，为连续型变量中的非数字类型赋值，满足后续分析建模需求，节省数据准备时间。

未来，专病数据库还将根据临床需求扩充数据项覆盖范围，增加专病科研随访平台，提升 NLP 算法性能，在数据维度、时效性和准确性方面不断提升。同时，提升数据分析建模能力，借助患者全景、多模态数据，结合传统 logistic 回归分析、决策树分类、深度神经网络等人工智能分析方法，对各类数据进行相关性分析，建立智能疾病预测模型，辅助指导临床决策。将数据转换为科研成果，最终回归临床，指导实践，提升专科科研水平，完善治疗方案，为患者提供更加优质的服务。

参考文献

1 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.

2 俞鹏飞，罗颢文，刘建模，等．面向医院的大数据治理模型设计［J］．医学信息，2021，34（10）：18－20.

3 VOS T，LIM S S，ABBAFATI C，et al. Global burden of 369 diseases and injuries in 204 countries and territories，1990—2019：a systematic analysis for the global burden of disease study 2019［J］. Lancet，2020，396（10258）：12041222.

4 国家统计局．卫生［EB/OL］.［2022－09－15］. https：//data. stats. gov. cn/easyquery. htm? cn = C01.

5 国家统计局．人口［EB/OL］.［2022－09－15］. https：//data. stats. gov. cn/easyquery. htm? cn = C01.

6 姚辉，范凯婷，王冉，等．急性缺血性脑卒中患者超早期活动的研究进展［J］．神经疾病与精神卫生，2022，22（1）：64－67.

7 邓军增．医院健康医疗数据治理探讨［J］．医学信息学杂志，2021，42（8）：14－17.

8 BERECZKI D，MIHÁLKA L，FEKETE I，et al. The debrecen stroke database：demographic characteristics，risk factors，stroke severity and outcome in 8 088 consecutive hospitalised patients with acute cerebrovascular disease［J］. International journal of stroke，2009，4（5）：335－339.

9 PATIL N，ARORA S，DAVIS L，et al. Incidence and predictors of 30－day acute cerebrovascular accidents post atrial fibrillation catheter ablation（from the nationwide readmissions database）［J］. The American journal of cardiology，2021，138（1）：61－65.

10 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.

11 吕旭东，田琪，蔡海领．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.

12 李雪迎，沙若琪，姚晨，等．面向真实世界数据的临床研究数据治理模式选择［J］．中国循证医学杂志，2020，20（10）：1150－1156.

13 罗辉，薛万国，乔岫．大数据环境下医院科研专病数据库建设［J］．解放军医学院学报，2019，40（8）：713－718.

14 胡健平，张学高．医院数据治理框架、技术与实现［M］．北京：人民卫生出版社，2019.

15 李丹彤，梁会营，刘广建．临床科研数据库建设中的数据标准化问题探讨［J］．中国数字医学，2021，16（1）：29－34.

16 张正宇，于跃，周虎，等．基于 OMOP 通用数据模型的 FAERS 数据库标准化与数据挖掘［J］．山东农业大学学报（自然科学版），2019，50（3）：434－437.

17 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.

18 缪姝妹，王忠民，景慎旗，等．医院患者主索引系统的设计与探索［J］．中国数字医学，2016，11（7）：61－63，66.

19 王毅豪，尚诗，袁骏毅，等．基于企业级患者主索引构建高脂血症专病科研数据库研究［J］．中国医学装备，2022，19（7）：116－120.

20 曹晓均，韦晓燕，毛钤镶．医院专病数据治理实践［J］．中国数字医学，2021，16（11）：17－20.

《医学信息学杂志》版权声明

（1）作者所投稿件无“抄袭”“剽窃”“一稿两投或多投”等学术不端行为，对于署名无异议，不涉及保密与知识产权的侵权等问题，文责自负。对于因上述问题引起的一切法律纠纷，完全由全体署名作者负责，无需编辑部承担连带责任。（2）来稿刊用后，该稿包括印刷出版和电子出版在内的出版权、复制权、发行权、汇编权、翻译权及信息网络传播权已经转让给《医学信息学杂志》编辑部。除以纸载体形式出版外，本刊有权以光盘、网络期刊等其他方式刊登文稿，本刊已加入万方数据“数字化期刊群”、重庆维普“中文科技期刊数据库”、清华同方“中国期刊全文数据库”、中邮阅读网。（3）作者著作权使用费与本刊稿酬一次性给付，不再另行发放。作者如不同意文章入编，投稿时敬请说明。

《医学信息学杂志》编辑部