

基于数据治理的专病数据库建设实践

龙思哲 吴震天 黎鹏安 张敦明 王 毅 周 毅

(中山大学附属第一医院信息数据中心 广州 510080)

(中山大学中山医学院 广州 510080)

张武军

(中山大学附属第一医院信息数据中心 广州 510080)

〔摘要〕 提出运用数据治理理念和全生命周期技术路径解决专病数据库建设实践中的数据相关问题, 阐述专病数据治理方法, 介绍专病数据库核心功能, 加速真实世界临床研究的成果转化, 助力专病、学科高质量发展。

〔关键词〕 专病数据库; 数据治理; 数据标准; 自然语言处理; 数据结构化

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.07.004

Construction Practice of the Specialized Disease Database Based on Data Governance LONG Sizhe, WU Zhentian, LI Peng'an, ZHANG Dunming, WANG Yi, Center of Information Technology and Statistics, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China; ZHOU Yi, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China; ZHANG Wujun, Center of Information Technology and Statistics, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China

〔Abstract〕 The paper puts forward the application of data governance concept and full life cycle technology approach to solve the data related problems in the construction of the specialized disease database, expounds the methods of specialized disease data governance, and introduces the core functions of the specialized disease database, so as to accelerate the transformation of real-world clinical research results, and contribute to the high-quality development of specialized disease and discipline.

〔Keywords〕 specialized disease database; data governance; data standard; Natural Language Processing (NLP); data structuring

1 引言

2016 年 6 月国务院办公厅发布的《关于促进和规范健康医疗大数据应用发展的指导意见》提出, 健康医疗大数据是国家重要的基础性战略资源, 要大

力推动政府健康医疗信息系统和公众健康医疗数据互联互通、开放共享, 消除信息孤岛, 积极营造促进健康医疗大数据安全规范、创新发展的发展环境^[1]。2021 年 6 月发布的《关于推动公立医院高质量发展的意见》指出, 医院应加强临床专科建设, 以专科发展带动诊疗能力和水平提升; 推进医学技术创新, 加强基础和临床研究; 强化信息化支撑作用, 推动大数据等新一代信息技术与医疗服务深度融合^[2]。随着健康医疗大数据价值凸显, 基于真实世界数据的临床研究成为医学高质量发展新动力^[3]。医院信息系统

〔修回日期〕 2022-06-02

〔作者简介〕 龙思哲, 硕士, 副主任技师, 发表论文 20 余篇; 通信作者: 张武军, 硕士, 副研究员。

是专科、专病数据的重要来源。以往系统建设以业务流程为导向,信息标准化程度、互联互通成熟度不足,产出数据质量难以保证,给专病数据库的建设和临床研究带来巨大挑战^[4-5]。中山大学附属第一医院(以下简称中山一院)是国家卫生健康委员会部属大型三甲综合医院,2017 年建设的基于卫生信息交换平台(Healthcare Information Exchange, HIE)^[6]的临床数据中心(Clinical Data Repository, CDR),已汇入 2003 年至今约 155 万份住院病历,包括住院病案首页、电子病历文书、检查、检验、病理、医嘱、护理记录、用血记录等。2020 年起启动专病数据库建设,通过全周期治理,逐步解决数据标准不统一、采集低效、精准度不足、质量无法核查和溯源等问题,形成长期持续、动态实时、初具规模、日趋完善的数据资产积累,构建统一化、标准化、结构化、可视化的专病数据库,加速临床病例资源到医学研究资产转化^[7-8]。

2 专病数据治理

2.1 医学术语标准化

目前国内尚无统一的医学术语标准,且部分标准存在缺乏整体概念体系、结构化定义不清、覆盖面不全等问题,为建立与需求相适应的标准化归集体系,专病数据库集成医学系统命名法-临床术语(Systematized Nomenclature of Medicine Clinical Terms, SNOMED CT),观测指标标识符逻辑命名与编码系统(Logical Observation Identifiers Names and Codes, LOINC (R)),中文人类表型标准用语(The Chinese Human Phenotype Ontology, CHPO), RxNorm, 国际疾病分类第 9 次、第 10 次修订本(International Classification of Disease V9, ICD-9/10)等数据标准^[9],定义临床术语本体中的重要概念类、类之间关系、类属性及属性分面。建立术语间关系,在相同条件下具有同义关系的术语赋予唯一代码,实现多源数据互联互通。

2.2 数据采集与交互

2.2.1 抽取-转换-加载(Extract Transform Load, ETL) 应用 ETL 技术快速接收业务

系统大量数据,减少数据采集对生产业务系统的影响,实现多个数据源的同时统一采集,避免数据转换系统对数据源的直接操作,同时进行清洗和标准化操作,包括数据字段映射、字典转换、跨域主索引患者主索引(Enterprise Master Patient Index, EM-PI)、数据校验等步骤。

2.2.2 交互管理 表视图查询方式便于接口开通和维护,数据入库效率高,平台主要通过其实现与院内生产系统的数据对接;Web Service 对接方式较固化,传输安全性能较高,更多应用于隐秘性较强的数据对接通道构建;部分数据可通过备份库在业务数据仓库(Operational Data Store, ODS)直接还原,适用于获取历史数据,但增量数据同步效率受限。综合考虑时效性、安全性、统一管理等因素,目前主要采用 Web Service、消息推送进行不同类型数据的交互。

2.2.3 同步机制 数据同步时间一般根据需求分 T+0、T+1、T+N 等,不同模块、类型的增量数据对接方式、对接频率可差异化处理,如患者基线数据、基础诊疗信息实时推送;检查、检验、病理数据由于存在滞后或延时,考虑在实时推送基础上增加定期遍历搜索前一时段内该部分数据的功能,保证患者信息完整性。

2.3 数据集成

专病原始数据来自多个信息系统,包括医院信息系统(Hospital Information System, HIS)、检验信息系统(Laboratory Information System, LIS)、医学影像存储与传输系统(Picture Archiving and Communication System, PACS)、病案、电子病历、手术麻醉等业务数据。通过梳理患者全诊疗过程产生的业务数据,形成 15 个标准的业务域分类和 221 个标准的检索项,见图 1。标准业务域主要包括人口学信息、就诊信息、诊断信息、一般检验、微生物检验、检查报告、医嘱记录、中医药方、门急诊病历、住院所有病历、入院记录、出院记录、手术记录、病案手术、生命体征,还可根据需求进行扩展,新增重症监护、肿瘤放化疗、生物样本库、生物信息多组学、随访队列等^[10]。

人口学信息	就诊信息	疾病分型	病史特点	体格检查
✓ 姓名 ✓ 性别 ✓ 出生日期 ✓ ABO血型 ✓ 现住址 ✓ 职业 ✓	✓ 就诊时间 ✓ 住院天数 ✓ 入院途径 ✓ 主要诊断 ✓ 住院费用 ✓	✓ 临床分型 ✓ 病理分期 ✓ 活动度 ✓ 肠外表现 ✓ 蒙特利尔分型 ✓	✓ 起病时间 ✓ 症状表现 ✓ 既往疾病 ✓ 家族史 ✓ 过敏史 ✓ 用药史 ✓	✓ 身高 (cm) ✓ 体重 (kg) ✓ 体温 ✓ 脉搏 ✓ 腹部压痛 ✓ 肛门病变 ✓
实验室检查	评分	影像学检查	内镜检查	病理检查
✓ 血常规 ✓ ESR ✓ 生化全项 ✓ C-反应蛋白 ✓ 结核试验 ✓ EB病毒抗体 ✓	✓ CDAI评分 ✓ Mayo评分 ✓ CDEIS评分 ✓ 组织学评分 ✓ 影像学评分 ✓	✓ CTE ✓ CT ✓ 超声 ✓ MRE/MR ✓ 造影 ✓ 心电图	✓ 病变性质 ✓ 累及范围 ✓ 溃疡形状 ✓ 溃疡大小 ✓ 溃疡面积	✓ 内镜小标本 ✓ 手术大标本 ✓
药物治疗	手术治疗	出院情况	随访信息	其他
✓ 氨基水杨酸类 ✓ 糖皮质激素类 ✓ 免疫抑制剂类 ✓ 营养治疗类 ✓ 生物制剂类 ✓	✓ 手术名称 ✓ 手术时间 ✓ 手术方式 ✓ 麻醉方式 ✓ 手术时长 ✓	✓ 出院日期 ✓ 出院诊断 ✓ 出院带药 ✓ 出院时情况 ✓	✓ CDAI评分 ✓ 症状 ✓ 体征 ✓ 检验检查 ✓ 药物浓度监测 ✓ 治疗方案调整 ✓	✓

图 1 专病通用数据集

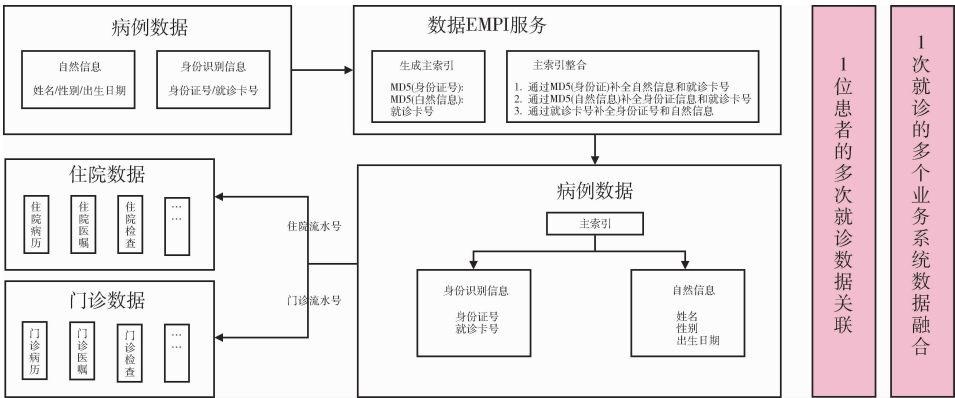


图 2 患者主索引设置

2.4.2 诊疗流程融合 梳理专病数据集 - 诊疗活动、时序逻辑 - 数据采集来源的对应关系，以关键事件的发生时间为标志，形成患者诊疗时间轴，实现数据逻辑有序排列。以特定展示、查询方式回嵌业务系统，用于复诊、随访等业务场景，见图 3。

由于不同数据源的时间跨度不同，不同事件之间有复杂的时序逻辑，需分层次整合不同时间颗粒度的数据，利于在后续分析中，通过特点指标回顾或预测患者转归或预后，为优化专病诊疗路径提供依据^[12-14]。

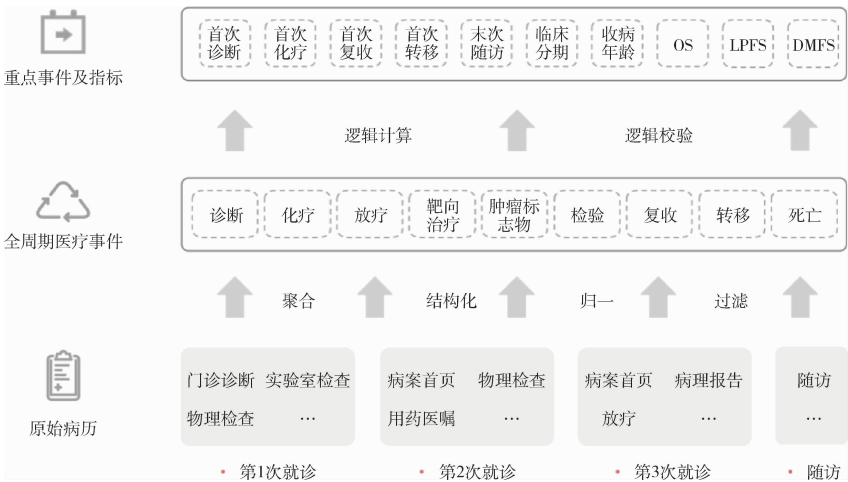


图 3 专病诊疗流程融合（以肿瘤为例）

2.5 数据归一

参照诊断、手术、药品、检查检验项目等数据标准和项目编码目录，通过自然语言的同义词表、医学术语的同义关联词表，在数据挖掘算法指导下，对文字表达不同但含义相对一致的字段信息进行归一，为后续和上层应用提供正确且统一的信息表达。以标准表为基准，将数据实际出现的医疗词向标准表做映射，包括机器学习加人工标注的过程。当出现标准表难以覆盖实际数据时，由专业医师决定是否扩充标准表。

2.6 数据结构化

2.6.1 后结构化 针对历史文本类数据，运用自然语言处理（Natural Language Processing, NLP）技术分析文本隐含语义和上下文结构关系^[15]，发现和挖掘结构化所需的知识模块，主要包括分词、同义词、主题和属性词关系图谱、上下文匹配逻辑、规则库、正则模板等^[16-17]。首先，进行专病文本自主学习、自动标注，计算准确率、召回率；其次，自动标注不理想的数据元，筛选由近及远的标准病例，由临床医师人工标注；再次，进行机器学习并建模，增加样本量迭代完善标注模型；最后，标注结果人工审核，不断迭代完善，满足专病不同语料数据要求^[18]。

2.6.2 前结构化 一般规律、有效信息的密度越高数据前结构化的价值越高、应用效果越好。主要针对病历文书、医技检查报告、功能评分量表等文本的增量数据，进行前瞻性结构化采集、存储和预处理，见表 1。前结构化改造由专病科室发起，涉及检查报告的还需联合医技科室共同设计模板、数据采集、运算逻辑、触发场景等，满足临床科研数据精准、精细的应用要求。

表 1 前结构化改造范畴

模板来源	模板类型（部分举例）
病历文书	入院记录、出院小结、首程记录、手术记录
放射检查报告	DR、CT、MR、PET-CT、PET-MR 等
超声检查报告	心脏、腹部、浅表、妇产等
内镜检查报告	肠镜、胃镜等
病理报告	活组织检查、手术小标本、手术大标本
功能评估量表	功能评分表等

2.7 数据质量核查

通过优化业务流程、完善接口性能，保障数据完整性；设置数据录入规则、逻辑校验机制等保障准确性；以患者号（住院号、门诊卡号）为第 1 维度，就诊流水号为第 2 维度，强化业务数据关联性；核查字典值域范畴，完善字典版本管理，约定业务数据流向，保障数据一致性；判断时间逻辑关系、同步机制，保障及时性。此外还有规范性、唯一性等控制指标，缺失值、无效值、重复值等异常发现与处理。

2.8 数据安全管控

设置权限管理模块，根据合法合规、最小需求原则为用户分配不同的角色、用户组，防止用户越级操作造成数据泄露；对信息更改等重要操作进行日志自动记录，提供用户名、访问记录、IP、登录时间、操作内容等用于回溯；针对敏感级别较高的数据采用加密或安全传输协议进行处理，对敏感字段的操作应有日志或报表供审核，严格控制对敏感数据的复制、导出等功能^[19]。

3 专病数据库核心功能

3.1 科研中心

根据研究目的创建科研项目，自定义研究队列纳入和排除条件；根据条件进行检索，利用原文检索、高级检索及条件树检索进行模糊检索或精确查询；根据检索到的病例对接院内患者全息展示，基于结构化内容进行视图展示，核查圈定研究对象；从病种数据映射，自定义衍生变量，形成所需的临床实验观察表（Case Report Form, CRF）数据项变量；进一步核查所筛选的数据是否符合要求，经必要的审核流程后，以多种形式导出，供后续统计分析^[20]。

3.2 病例中心

支持按照患者的门诊、住院就诊时序逐个展示单个就诊次序下的患者全量诊疗信息，内容包括基本概

览、病历、医嘱、检验、检查、量表、护理等信息。展示患者核心指标信息及对应的疾病信息，实现患者个人画像视图构建，方便临床科研人员更直接查看患者总体诊疗情况，辅助临床科研入组判断。

3.3 随访中心

根据随访目的创建随访项目，自定义队列纳入

和排除条件；根据患者身份证或就诊卡号进行检索，识别目标随访用户；通过签署知情同意书、授权书等将患者纳入随访队列进行后续随访计划安排；根据随访队列的研究需求制定随访表单，并对随访进行预警分析和干预；制定失访等特殊事件的管理策略和处理流程，随访完成的问卷数据按权限申请导出下载，见表 2。

表 2 专病数据库通用功能

1 级功能	2 级功能	功能列表	功能描述
首页	专病概览	数据中心分布	病例分布、核心指标、数据完整度等
科研中心	科研项目管理	项目列表	所有项目列表，可新增、编辑和删除项目。可根据数据权限展示全院/所属科室项目
		项目信息	可查看及维护项目基本信息，展示项目数据、项目结题及说明
		操作流程	查看科研流程指引，可按照流程跳转至对应操作页面
	病例	病例列表、自定义	队列/亚队列病例管理，可提供手工批量导入/检索/从上级队列入组病例；病例标签
		列、搜索菜单、标签维护	管理；病例随访操作；病例文件管理；病例状态记录
	表单	科研 CRF 随访问卷表单编辑	项目 CRF/问卷管理，提供选择已映射数据元和创建衍生变量；可视化 CRF/问卷设计器，拖拽生成 CRF 和题型；提供逻辑跳题、多维度评分等高级设置
	随访	评估问卷设置、知情同意书设置、随访计划设置、问卷分析、随访分析	队列随访管理，可设置随访流程，包括评估设置、知情同意书设置、随访计划设置、启动/暂停随访。提供随访相关数据分析，包括风险预警、随访分析、问卷分析
	病例中心	患者全息视图	患者基本信息业务域、医嘱/处方业务域、检验、检查、手术、护理、分娩、病历文书、随访业务域展示
		患者时间轴视图	就诊时间轴、体征时间轴、检查时间轴、药品时间轴、检验时间轴展示
		患者画像	提取后结构化指标，展示患者行为数据
		结构化数据项视图	患者基本信息业务域、医嘱/处方业务域、检验、检查、手术、护理、病历文书、随访业务域结构化展示
		原文检索	原文任意搜索，可按照就诊维度/患者维度检索
		条件树检索	基于病历节点与或非逻辑的全结构化数据项检索，可分组检索，可按照就诊维度/患者维度检索
		高级检索	基于病历节点与或非逻辑的检索，可按照就诊维度/患者维度检索
		语义化检索	阳性症状语义化检索
		录入病例	可新增病例、维护病例，建立患者主索引
			录入单个病例的问卷、量表，关联患者
			按照固定病历模板录入单个患者的病历，关联患者具体某次就诊
随访中心	问卷管理	问卷管理	可选择问卷模板生成表单、自定义生成表单，可维护表单，生成表单二维码供患者扫码填写
		待办列表	到达今日随访点的待随访患者及超期未完成患者列表
		全息病历	查看患者全息病历
		随访详情	可标记患者、查看患者报告、填写患者 CRF
	异常列表	提醒患者	提醒患者填写问卷
		标注异常列表	已标注随访异常的患者列表，可查看患者全息、患者详情，对异常患者进行恢复、中止操作

4 结语

截至 2021 年末中山一院已开展 20 个专病数据库建设, 入库病例数超过 50 万份, 60 多位专病医师、研究助理通过专病数据库进行真实世界数据探索, 已支撑多项回顾性、前瞻性临床研究, 将推动科学研究和临床工作的规范化、标准化、规模化, 进一步提升专病临床研究能力, 反哺和促进学科发展。通过专病数据库建设实践, 发现业务系统、集成平台在数据标准、交互机制等方面存在的不足, 需加强临床科室、管理部门、信息部门、承建商的高效沟通协作。下一步将继续以专病需求为导向、问题为切入点, 推进数据治理的标准化、规范化、合理化落地, 逐步构建基于专病数据资产的综合治理体系。对数据库功能进行迭代和扩展, 形成院前-院中-院后的专病全程闭环管理; 建立院级科研数据中心, 依托多中心共享平台完善专病区域联盟的数据资源合规共享和多元应用, 共同构建我国疾病诊疗规范。

参考文献

- 1 国务院办公厅. 关于促进和规范健康医疗大数据应用发展的指导意见 [EB/OL]. [2021-06-24]. http://www.gov.cn/xinwen/2016-06/24/content_5085211.htm.
- 2 国务院办公厅. 关于推动公立医院高质量发展的意见 [EB/OL]. [2021-11-10]. http://www.gov.cn/zhengce/content/2021-06/04/content_5615473.htm.
- 3 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.
- 4 曹晓均, 韦晓燕, 毛铃饕. 医院专病数据治理实践 [J]. 中国数字医学, 2021, 16 (11): 17-20.
- 5 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.
- 6 王觅也, 郑涛, 李楠, 等. 医疗大数据集成及应用平台体系构建 [J]. 医学信息学杂志, 2019, 40 (8): 37-42.
- 7 Davis S, Ehwerhemuepha L, Feaster W, et al. Standardized Health Data and Research Exchange (SHaRE): Promoting a

Learning Health System [J]. JAMIA Open, 2022, 5 (1): 1-7.

- 8 姜伟, 于天贵, 苟延农, 等. 健康医疗大数据信息资源目录体系构建 [J]. 医学信息学杂志, 2021, 42 (3): 20-23.
- 9 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.
- 10 陈嘉旂, 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.
- 11 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.
- 12 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.
- 13 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.
- 14 Baxter S L, Saseendrakumar B R, Paul P, et al. Predictive Analytics for Glaucoma Using Data from the All of Us Research Program [J]. American Journal of Ophthalmology, 2021 (227): 74-86.
- 15 Wang S Y, Pershing S, Tran E, et al. Automated Extraction of Ophthalmic Surgery Outcomes from the Electronic Health Record [J]. International Journal of Medical Informatics, 2020 (133): 1-6.
- 16 Bozkurt S, Park J I, Kan K M, et al. An Automated Feature Engineering for Digital Rectal Examination Documentation using Natural Language Processing [EB/OL]. [2021-11-10]. <https://pubmed.ncbi.nlm.nih.gov/30815067>.
- 17 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.
- 18 Seneviratne M G, Seto T, Blayney D W, et al. Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer [J]. Generating Evidence & Methods to Improve Patient Outcomes, 2018, 6 (1): 1-7.
- 19 弓孟春, 刘莉, 彭郁, 等. 新发重大传染病流行期间临床研究数据治理: 挑战与应对策略 [J]. 医学信息学杂志, 2020, 41 (8): 2-7.
- 20 戴静娟, 胡萍, 陈名扬, 等. 耳鼻喉科专病大数据平台的设计与实践 [J]. 中国卫生信息管理杂志, 2021, 18 (3): 405-410, 425.