膀胱癌专病科研数据库建设与应用

倪文丽[1],唐慧琴[1],雷铭德[1],史江华[1],胡海龙[1,2]

(天津市医科大学第二医院泌尿外科研究所[1],泌尿外科[2],天津 300202)

摘要:基于医院信息化平台建设膀胱癌专病科研数据库,通过自然语言处理技术实现专病临床数据的结构化和标准化,构建一个符合临床和科研的高质量、一体化的膀胱癌专病数据库。该数据库可以实现智能病例检索、患者科研全景、科研项目管理、定制化应用等功能,有效支撑医生的临床研究工作。本文针对该数据库的系统架构、数据集、数据整合及其应用情况进行了总结。

关键词:膀胱癌;专病数据库;自然语言处理;疾病数据集

中图分类号:R737.14 **文献标识码**:B **DOI**:10.3969/j.issn.1006-1959.2023.07.006

文章编号:1006-1959(2023)07-0031-05

Construction and Application of Scientific Research Database for Bladder Cancer

NI Wen-li[1],TANG Hui-qin[1],LEI Ming-de[1],SHI Jiang-hua[1],HU Hai-long[1,2]

(Institute of Urology[1],Department of Urology[2],the Second Hospital of Tianjin Medical University,Tianjin 300202,China)

Abstract:To build a scientific research database of bladder cancer based on hospital information platform, and realize the structure and standardization of specific clinical data through natural language processing technology, so as to construct a high-quality and integrated bladder cancer specific disease database in line with clinical and scientific research. The database can realize intelligent case retrieval, patient research panorama, scientific research project management, customized application and other functions, and effectively support the clinical research work of doctors. This paper summarizes the system architecture, data sets, data integration and application of the database.

Key words:Bladder cancer;Disease-specific database;Natural language processing;Disease dataset

膀胱癌(bladder cancer)为泌尿系统最常见的恶性肿瘤，每年约有 573 000 例新病例和 213 000 例死亡。男性比女性更常见,男性的发病率和死亡率分别为 9.5/10 万和 3.3/10 万，约为全球女性的 4 倍;是男性第 6 位最常见的癌症和第 9 大癌症死亡原因。2020 年,膀胱癌位居全球肿瘤新发病例数第 10 位,我国肿瘤新发病例数第 13 位[1]。2016 年我国膀胱癌新发患者人数为 7.7 万人,到 2020 年增长到 8.6 万人,复合年增长率为 2.68%,预计 2025 年将超过 10 万人[2]。为提高膀胱癌患者生存率和生存质量,在临床诊疗的同时需要开展大量临床研究。但科研数据采集方面仍存在一定困难,人工填报而并非客观获取,不仅增加医院的劳动负荷,而且数据质量低,为此,通过信息技术手段为临床科研提供数据支撑显得至关重要[3-5]。本研究以天津医科大学第二医院为背景,构建标准化、高质量的膀胱癌专病数据库,旨在为科研人员提供丰富的临床及科研数据,同时满足临床多样化的科研数据的采集。

1 膀胱癌专病库的系统架构

膀胱癌专病数据库依托于临床数据中心进行建设,以膀胱癌病患者为研究对象。专病库系统架构从下到上分共为 5 层,分别是数据采集层、数据互通互联层、数据汇集层、数据分析加工层和数据应用层。为了实现不同系统间异构数据源的采集、存储、加工,且保证临床业务系统自身的稳定和独立,本系统采用分布式(Service Orieted Architecture,SOA)架构。s1 数据采集层中包括医院信息系统(Hospital Information System,HIS)、电子病历系统(Electronic Medical Record,EMR)等临床业务系统。s2 数据互通互联采用企业服务中心(Eterprise Service Bus,ESB)方式与各业务系统对接,既兼容了多源异构数据的采集,又保障了各系统自身的稳定性、安全性。s3 数据汇集层集中了院内临床数据和院外数据的合集。s4 数据分析加工层则是对采集到的数据进行二次加工,利用患者主索引、自然语言处理(Natural Language Processing,NLP)技术,形成结构化、归一化的数据集,生成高质量、标准化的膀胱癌专病数据中心。s5 数据应用层则根据不同的业务需要,主要包括搜索应用、科研应用、患者应用及医院管理应用等业务场景,提供了多样化的技术与数据支撑。膀胱

作者简介:倪文丽(1973.5-),女,天津人,本科,图书助理馆员,主要从事医院生物样本库信息管理工作

通讯作者:胡海龙(1978.1-),男,黑龙江大庆人,博士,主任医师,教授,博士生导师,主要从事膀胱癌的个体化及微创治疗研究

癌专病数据库以行业的标准规范为指导，利用多源异构数据采集、分布式存储、数据抽取、转换、加工等信息技术手段，统一集成膀胱癌患者的临床数据样本，建立结构化、标准化的专病数据库，实现专科数据的互联互通。膀胱癌专病库系统架构设计，见图 1。

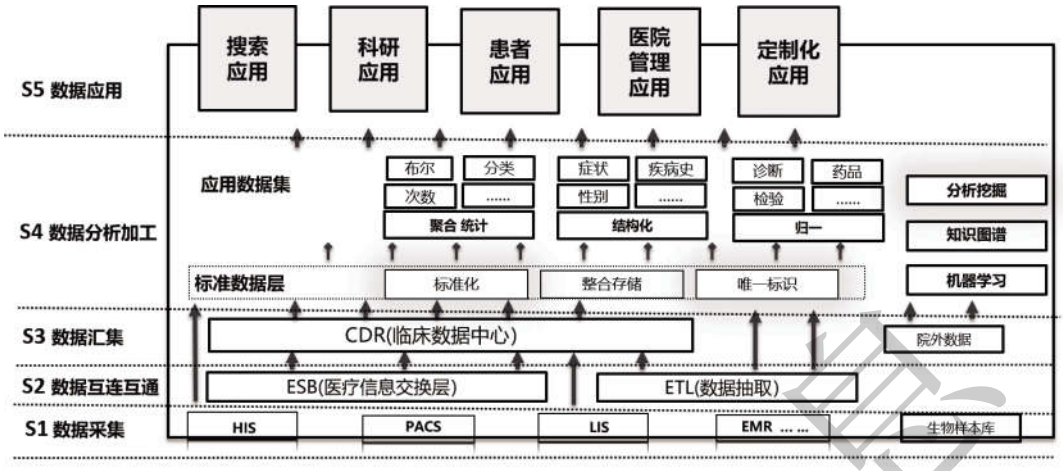

图 1 膀胱癌专病库系统架构设计

2 膀胱癌专病库数据集

传统科研数据的采集主要以研究目的为主导，大多采用自定义形式构建模型抽取数据，导致数据标准的不统一，在后续开展多中心研究时容易产生共享障碍。为解决这一问题，本研究参考了国内外行业标准和规范，通过建立标准化、动态化的膀胱癌数据集，使临床数据在存储、交换过程中遵循统一标准，保证数据的可比性和一致性。膀胱癌数据集由通用数据集和疾病特征数据集组成，通用数据集是指适用于泌尿外科疾病的数据项，疾病特征数据集则用于描述膀胱癌专病所独有的数据项。这种构建方式不仅便于对其它泌尿外科病种的横向拓展，同时也支持对膀胱癌专病数据维度的纵向延伸[6]。

数据库标准化建设是卫生健康信息标准化建设的长效机制，以国际标准中国化、中国标准国际化为方向[7,8]。因此在制定通用数据集时，参考国家卫生健康委行业数据标准，中国卫生信息数据元值域代码 ws364.x-2011、电子病历基本数据集 ws445.x-2014；相关术语标准参考国际疾病分类（第 10 版）ICD-10 和国际疾病分类(第 9 版)临床修订第三卷：手术与操作 ICD-9-CM-3；国家相关数据标准参考 GB/T 2269.9-2003 个人基本信息分类与代码第 9 部分人的性别分类和 GB/T 4671-2008 家庭关系代码等。构建膀胱癌疾病特征数据集时，以相关疾病诊疗指南为指导[9]，如 AJCC/UICC 临床分期手册、EAU 指南（肌层浸润和转移性膀胱癌）、JUA 临床实践指南（膀胱癌）、中华医学会(膀胱癌治疗指南)等[10-13]。膀胱癌通用数据集与疾病特征数据集两者相结合，汇总构建成膀胱癌专病数据集，包含 16 个模块、648 个数据项，构成膀胱癌专病数据库的建设。膀胱癌专病数据集，见表 1。

表 1 膀胱癌专病数据集

分类	模块名称	模块内容
通用数据集	患者人口学信息	姓名、性别、出生日期、民族、职业类型等
	就诊记录	住院号、入院科室、入院日期、主诊断等
	一诉五史	主诉现病史、既往史、个人史、家族史等
	体格检查	入院体重、身体状况评价等
	诊断	诊断时间、诊断名称、诊断来源等
	检查	超声检查、CT 检查、MRI 检查等

表 1(续)

分类	模块名称	模块内容
通用数据集	分子免疫标志物	记录时间、免疫组化、基因检测信息等
	检验	血常规、尿液分析、生化检查等
	放射治疗	开始时间、医嘱数量、频次等
	肿瘤药物治疗	开始时间、结束时间、药物名称、用药频次等
	医嘱	医嘱项目类型、开立科室、药物剂型等
疾病特征数据集	尿路造影	膀胱内是否见肿物、是否上尿路梗阻等
	专科检查	腹部外形、凹陷部位、是否触及腹部肿块等
	病理	脱落细胞学检查、膀胱病理信息、后尿道病理信息等
	手术治疗	术前诊断、TURBT、膀胱镜治疗集探查等
	诊疗概览	是否行膀胱癌根治术、是否行尿道膀胱肿物切除术等

3 膀胱癌专病数据整合

数据整理是在挖掘提炼数据价值的过程中进行的前期的数据预处理工作，数据整理是数据整合的基础，并相辅相成[14]。医院各个信息系统存在“数据孤岛”、深度数据治理高度依赖医生手工操作、数据难以共享复用、复杂临床事件难以有序组织排列等问题[15]。使得医疗数据无法有效地整合利用，从而导致真实世界大数据研究无法开展或效率低下[16]。通过科研大数据平台的数据库同步技术和数据仓库技术（Extract-Transform-Load，ETL）等，对院内各个业务系统的数据，主要包括 HIS、实验室信息（LIS）病理等进行同步、抽取，实现针对医院多个信息系统的多源异构数据的采集和汇聚，确保符合条件的病例自动、持续性入库（膀胱癌全库纳入条件为全部诊断有描述膀胱/尿路上皮/脐尿管癌/瘤/CA/MT，且中间不含“良性”，或诊断 ICD10 编码包含 C7），形成全量、连续、完整、可再利用的数据资产。整合医院信息系统多数据源构建专病数据库，见图 2。

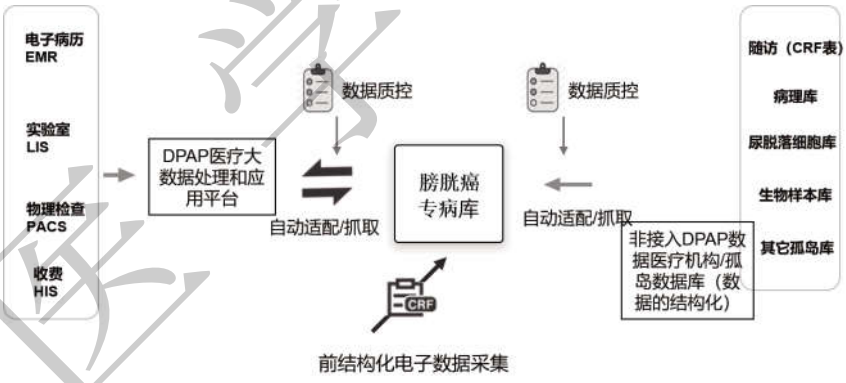

图 2 整合医院信息系统多数据源构建专病数据库

在临床数据中，根据数据的性质、来源、质量的不同可分为两个级别：第一级的数据以结构化存储、质量较高为特点，用于记录患者的基本情况、诊断、医嘱等重要信息，性别、年龄等人口学信息以及药品医嘱等，通过建立规范的映射存储关系进行抽取利用。第二级的数据以长文本为主的非结构化数据，如现病史、既往史等，通过自然语言处理技术将大段自然语言转化为标准字段和阈值，满足科研数据分析需求用于统计分析[17]。

4 膀胱癌专病库应用

本院自 2017 年开始建设基于医院数据平台的膀胱癌单病种数据库，截至 2022 年 3 月，全库已经纳入自 2005 年 3 月起在本院就诊的患者 33 692 例。4.1 患者病历检索 专病库提供对患者的病历检索功能，包括关键词搜索、条件树搜索、精准搜索 3 种方式，关键字搜索从全部病历中进行匹配，适合数据探查等。条件树搜索适用于患者或者病历维度多条件匹配与时间无关的场景。精准搜索适用于需要对

患者或者病历在时间线上进行多个条件的逻辑判断的场景。3 种场景的搜索极大节约了医生时间，更快速地定位到符合研究要求的患者，将患者历次就诊以结构化数据的明细结果导出，便于对数据进一步综合分析和使用，提高了科研产出效率。

4.2 患者时间轴 以关键诊疗事件的发生时间为标志的患者全周期时间轴可视化，实现数据的逻辑有序排列，从而更好地支持临床及科学研究[18,19]。患者的病历数据可以看作一条时间序列，记录着患者历次就诊的病史采集、病情分析、诊断、治疗的临床全过程。在这条时间序列轴上，症状、体征、化验和检查指标、并发症/合并疾病、用药和手术等可以视作诊疗的关键节点，用于后续的数据挖掘。不同数据源的时间跨度不一致、不同事件之间有复杂的时序依赖关系。整合数据有助于查看特定患者重点临床事件的时序关系。通过梳理患者重点事件数据的全周期时间轴，能够对患者诊疗事件按照时间进行可视化展示和分析，直观了解患者重点诊疗事件和结局，以及重点指标的时序进展情况，例如重点诊疗事件（如：首次诊断、主要症状、首诊病理分型等）、重点诊疗事件的发生时间（如：首诊时间、首次手术时间等）、同一时间轴跨度下查看多个重点指标的进展情况，见图 3。

图 3 重点事件数据的全周期时间轴

4.3 科研应用 基于膀胱癌专病库开展了多项膀胱癌的临床实验。因为每个临床实验的研究对象、数据处理、随访计划、项目进度及任务列表不尽相同，所以可以定制 CRF 表单个性化模板，进行模块化管理，现多个临床实验项目正在开展中。如卡介苗膀胱灌注在预防中、高危非肌层浸润性膀胱癌术后复发的有效性、安全性，及不同给药方案的随机、对照、多中心临床试验等。

4.4 知识全库 知识全库分由文献、指南共识、临床路径、药品说明书、临床试验、误诊误治模块构成。文献模块可以根据标题，作者等搜索关键词进行文献搜索，也可以进行高级检索。最新文献显示了最新的中英文文献。文献热点部分则显示膀胱癌的相关文献及其类型，如诊疗指南、Meta 分析、病历报告、临床研究、综述、系统评价等每种文献类型的数量。临床试验模块提供了在 ClinicaTrials 注册的临床试验的最新更新时间，招募状态等。指南共识和临床路径模块介绍了膀胱癌相关疾病的指南共识和临床路径。知识全库“一站式”满足临床医生的相关需求，从中提取更精准的专病信息，在诊断、治疗、康复各阶段辅助临床，提升研究效率。

5 总结

膀胱癌专病数据库的建设，提高了真实世界大数据研究质量与效率，一方面可以赋能临床研究、挖掘临床规律，另一方面能够帮助医生总结经验、提升疾病的诊治水平，为临床实践提供强有力的数据支撑。膀胱癌专病数据库建设将高效推动医院在膀胱癌领域的科学研究和临床工作。建设规范化、标准化、规模化的膀胱癌专病科研数据库，可以进一步提升膀胱癌临床研究的能力，加速成果转化。未来可望通过在区域乃至全国范围内推广专病数据库标准、建设流程规范，将分散于不同医院、不同信息系统中

的临床信息通过数据采集、存储、整合和挖掘等步骤集成云端的数据中心，对膀胱癌等肿瘤疾病的医疗数据进行规范集成、深度挖掘、综合利用，为后续开展多项真实世界多中心的研究提供强有力的保障。

参考文献：

[1]Sung H,Ferlay J,Siegel RL,et al.Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries [J].CA Cancer J Clin,2021,71(3):209-249.

[2]李辉章,郑荣寿,杜灵彬,等.中国膀胱癌流行现状与趋势分析[J].中华肿瘤杂志,2021,43(3):293-298.

[3]Okorie CL,Gatsby E,Schroeck FR,et al.Using electronic health records to streamline provider recruitment for implementation science studies[J].PLoS One,2022,17(5):e0267915.

[4]Wu WT,Li YJ,Feng AZ,et al.Data mining in clinical big data: the frequently used databases, steps, and methodological models [J].Mil Med Res,2021,8(1):44.

[5]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.

[6]宓林晖,潘常青,袁骏毅,等.心衰专病数据库的建设与应用[J].微型电脑应用,2022,38(2):12-15.

[7]李岳峰,胡建平,庹兵兵,等.我国卫生健康信息标准建设成效与思考[J].中国卫生信息管理杂志,2021,18(3):324-329.

[8]El-Khayat YM,Forbes CS,Coghill JG.Guideline.gov: A Database of Clinical Specialty Guidelines [J].Med Ref Serv Q,2017,36(1):62-72.

[9]尚诗,袁骏毅,岑星星.基于 EMPI 心肌病专病数据库的构建[J].中国医疗设备,2022,37(6):115-118.

[10]靳英辉,曾宪涛.中国非肌层浸润性膀胱癌治疗与监测循证临床实践指南 (2018 年标准版)[J]. 现代泌尿外科杂志,2019,24(7):516-542.

[11]Flaig TW,Spiess PE,Agarwal N,et al.Bladder Cancer, Version 3.2020, NCCN Clinical Practice Guidelines in Oncology [J].J Natl Compr Canc Netw,2020,18(3):329-354.

[12]Witjes JA,Bruins HM,Cathomas R,et al.European Association of Urology Guidelines on Muscle-invasive and Metastatic Bladder Cancer: Summary of the 2020 Guidelines [J].Eur Urol,2021,79(1):82-104.

[13]Babjuk M,Burger M,Capoun O,et al.European Association of Urology Guidelines on Non-muscle-invasive Bladder Cancer (Ta,T1,and Carcinoma in Situ)[J].Eur Urol,2022,81(1):75-94.

[14]杜小勇,陈跃国,范举,等.数据整理——大数据治理的关键技术[J].大数据,2019,5(3):13-22.

[15]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.

[16]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.

[17]王昳佳.大数据时代背景下医院数据中心建设的相关思考[J].电脑知识与技术,2020,16(3):9-10.

[18]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.

[19]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.

收稿日期：2022-07-07；修回日期：2022-08-11

编辑/肖婷婷