

• 综述 •

DOI: 10.3724/BNSFC-2025-0002

构建罕见疾病高质量生物医学数据集： 以儿童小脑髓母细胞瘤为例

刘海龙^{1,2,3}马骋宇⁴姜涛^{1,2,5}李艺影^{1,5,6}陈瑞⁴田雨^{4*}

1. 首都医科大学附属北京天坛医院, 北京 100070

2. 国家神经疾病医学中心, 北京 100070

3. 首都医学科学创新中心, 北京 100069

4. 首都医科大学公共卫生学院, 北京 100069

5. 北京市神经外科研究所, 北京 100070

6. 神经系统肿瘤靶向药物与细胞治疗北京市工程研究中心, 北京 102600

[摘要] 罕见肿瘤虽发病率低,但种类繁多,总数庞大,整体占全部恶性肿瘤20%以上。然而,样本稀缺、研究资源分散等现实情况直接制约了罕见肿瘤研究,精准诊断与靶向治疗更加依赖开放的高质量数据库支持。近年随着基因组技术发展,部分罕见肿瘤研究取得了一定进展,但数据壁垒为罕见肿瘤的精准医学发展带来严峻挑战。本文以儿童小脑髓母细胞瘤为例回顾了罕见肿瘤生物医学数据库建设现状,剖析数据壁垒对研究生态的影响,并提出未来高质量生物医学数据库构建应对策略。

[关键词] 罕见肿瘤;肿瘤数据库;精准医学;人工智能;髓母细胞瘤

罕见肿瘤(Rare Cancer)通常指年发病率低于2.5~15/10万的恶性肿瘤,涵盖神经系统肿瘤、胰腺肿瘤、软组织肉瘤、神经内分泌肿瘤等多个种类^[1]。尽管每种罕见肿瘤发病率低,但其总体发生比例可达20%~25%^[2]。尤其儿童神经系统肿瘤,存在诊断困难、治疗选择有限、预后较差等问题^[3]。由于罕见肿瘤致病机制复杂、组织学多样、缺乏标准治疗手段,更加依赖跨学科、跨机构开放的数据协作研究。

高质量的生物数据库与生物样本库是罕见肿瘤研究的重要基础^[4]。生物数据库系统性地整合临床、影像、病理与组学数据,支撑分子机制探索、分型建模与靶向治疗研究。生物样本库提供稳定可复用的组织、血液、DNA/RNA等材料,是高通量测序、蛋白质组分析及临床前模型构建的物质基础。通过整合多中心前瞻性队列、临床-组学数据联动、高通量测序及AI辅助分析技

术,这些数据库有望支撑罕见肿瘤领域多种科学问题的解决^[5,6],如跨亚型分子分型、关键靶点识别、复发转移机制研究及新药筛选等。

然而,近年来国际数据共享环境发生剧变,美国国家卫生研究院(National Institutes of Health, U. S., NIH)等限制中国等研究者访问其部分数据库,首当其冲的是全球最大、最权威的公开癌症数据库之一——SEER数据库,相关限制可能将继续波及美国国家生物技术信息中心(National Center for Biotechnology Information, U. S., NCBI)旗下的众多平台数据库,如基因型与表型数据库(dbGaP)、癌症基因组图谱(TCGA)、大规模基因数据分析的AnVIL平台等^[7-9]。这些数据库代表性强、数据量大、分析工具完善,存储了全球最核心的人类基因组、表型组信息和疾病研究数据,是解析癌症等慢性病、罕见病等复杂疾病的关键工具,是全球研究人员长期高

收稿日期:2025-05-26; 修回日期:2025-07-16

* 通信作者, Email: yutian@ccmu.edu.cn

本文受到国家科技重大专项青年科学家项目(2022ZD0210100)、中国工程院科技战略咨询项目(2025-XBZD-10-01)、国家自然科学基金项目(82273343, 82204127, 82172608)、北京市杰出青年科学基金项目(24JQG011)的资助。

引用格式: 刘海龙, 马骋宇, 姜涛, 等. 构建罕见疾病高质量生物医学数据集: 以儿童小脑髓母细胞瘤为例. 中国科学基金, 2025, 39(4): 635-642.

Liu HL, Ma CY, Jiang T, et al. Establishing a high-quality biomedical resource for rare diseases: Insights from pediatric medulloblastoma. Bulletin of National Natural Science Foundation of China, 2025, 39(4): 635-642. (in Chinese)

度依赖的数据库。对于样本稀缺、研究资源高度分散的罕见肿瘤而言,国际数据库的准入限制将直接影响精准诊断和靶向治疗的进展。

近年来,我国已启动罕见病信息平台、“天枢计划”和国家基因组科学数据中心等一系列数据库建设工作,初步形成涵盖部分罕见病种的分布式数据体系^[10]。然而,当前我国罕见肿瘤数据库仍存在病种覆盖率不足、数据标准不统一、跨机构共享机制缺失等问题,难以满足高质量科研与转化医学发展的需求。为此,本文以儿童小脑髓母细胞瘤为代表性病例,系统梳理罕见肿瘤生物医学数据库建设现状,探讨罕见肿瘤高质量数据库构建的路径与策略,以期对相关数据库体系建设提供可借鉴的实践参考。

1 罕见肿瘤研究的核心挑战:数据资源稀缺

1.1 样本稀缺与专病队列构建困难

罕见肿瘤的定义因地区而异,美国国家癌症研究所(National Cancer Institute, U. S., NCI)将年发病率低于15/10万的癌症定义为罕见癌症^[11],欧洲罕见癌症监测项目采用6/10万的阈值^[12],而中国国家癌症中心登记处则使用更严格的2.5/10万标准^[13]。尽管该比例相对较低,但考虑到中国庞大的人口基数,罕见肿瘤患者的绝对数量仍然相当可观,然而样本分布分散、难以集中成为研究的首要障碍,这一问题在美国关闭部分数据库后变得更加严峻。

样本分散性是构建高质量专病队列的主要瓶颈。以儿童小脑髓母细胞瘤为例,发病率小于3~5/10万,并且这些病例分散在全国数百家医院,单个一般医疗中心每年接诊量通常不足20例。我团队依托首都医科大学附属北京天坛医院虽建立了儿童髓母细胞瘤专病队列,目前已纳入1 024例具完整诊疗信息病例,其中70%具有高质量遗传学信息,但相较于国际儿童髓母细胞瘤联盟的数据体量,仍具有一定差距。同时,仅依赖单一中心的数据,难以满足深度研究的需求。

专病队列构建还面临标准化不足的挑战。中国各医疗中心在罕见肿瘤的数据采集方面尚缺乏统一标准,数据格式差异较大^[14,15]。生物样本库建设滞后也是制约因素,罕见肿瘤样本库多缺乏标准化管理,样本保存质量参差不齐,临床信息完整性较差,大多数罕见肿瘤样本因保存不当或信息缺失而无法用于深入研究。伦理与共享机制的缺失也进一步加剧了样本稀缺问题。当前中国患者数据跨机构共享需逐例签署知情同意,效率低下。相比之下,欧洲罕见癌症监测项目通过建立统一的数据共享协议和伦理豁免机制,有效提升了数据整

合能力^[16]。中国在罕见肿瘤样本共享方面缺乏国家级协调机制,各机构往往各自为政,造成资源浪费和研究重复。

1.2 制约精准诊断与治疗方案

罕见肿瘤的诊断本身就是一个巨大的临床挑战,而数据资源稀缺进一步提高了这一挑战的难度。美国核心生物医学数据库的关闭进一步限制了中国研究人员获取全球罕见肿瘤分子特征数据的能力,对构建精准诊断体系建立造成了显著障碍。罕见肿瘤需要综合诊断方法,包括临床评估、影像学检查、分子和病理学检测等。病理诊断作为罕见肿瘤识别的金标准,但由于其稀少性和组织学多样性,加之部分病理医生缺乏相应经验和专业知识,常导致诊断延迟^[17]。同时,分子诊断严重依赖大规模数据支持,故在数据资源受限的环境下可能进一步恶化。另外,罕见肿瘤的影像诊断同样受到数据资源限制的显著影响。影像学标准的建立依赖于大量案例数据形成影像组学的积累,我国虽然患病人数庞大,但因就诊分散、缺乏统一数据管理,目前缺乏足够的数据支持类似诊断体系的开发,也限制了人工智能辅助模型在该领域的训练^[18-20]。

罕见肿瘤缺乏标准治疗,一般常用放疗方案或化疗药物效果有限,常见肿瘤靶向药物对罕见肿瘤疗效也不确定,且在不同亚型间存在显著差异。罕见肿瘤治疗方案研发推进本就面临重重困难^[11],而数据资源稀缺更进一步放大了这一挑战^[21]。美国生物医学数据库的关闭使研究人员无法获取关键的临床试验数据、药物反应谱和治疗方案比较研究,严重制约了罕见肿瘤治疗领域的进展,并显著延长新药研发周期,增加研发成本,最终损害患者福祉。临床试验数据获取受限是治疗方案开发的首要障碍,信息不对称将直接影响中国研究人员设计临床试验的能力和患者获得最新治疗的机会^[22]。其次,罕见肿瘤通常表现出异于常见肿瘤的独特分子遗传特征,针对这类靶点的药物开发需要大量组学数据支撑,数据库的关闭使研究人员难以全面了解全球研究进展和失败教训,显著增加研发风险。另外,罕见肿瘤样本稀缺直接导致临床前模型匮乏,严重制约基础研究进展^[23]。患者来源异种异体移植瘤模型在罕见肿瘤中的构建成功率低,而类器官培养技术也面临巨大挑战。

2 高质量生物医学数据库构建:推动与桎梏并存

2.1 国际罕见肿瘤生物医学数据库建设现状

随着精准医学的推进,国际上逐步认识到构建系统化、可共享的罕见肿瘤数据库对于推动机制研究和临床转化的重要性,如以欧洲为代表的Rare Cancers

Europe项目^[24],以及美国NCI主导的Rare Tumors Initiative项目^[25]。此类数据库建设主要呈现以下几个特点:(1)多中心协作的网络化建设:多中心协作模式,整合不同区域的病例信息、生物样本、临床结局和组学数据;(2)强调临床—分子数据一体化:许多数据库注重将临床资料与分子组学数据(如基因组、转录组、表观组)深度整合;(3)开放共享与数据标准化:国际数据库普遍强调数据标准化和开放共享,以提升数据可重复性和跨研究机构可比性;(4)跨平台协同与人工智能赋能:随着人工智能技术的发展,多个数据库平台开始整合AI工具用于影像识别、突变预测、预后建模等,提升数据使用效率。

但是,“数据孤岛”现象仍然普遍存在,如区域性分散存储、机构间共享壁垒、数据标准不统一等。四川大学华西医院Xue等人^[26]的研究显示,中国罕见肿瘤研究“数据孤岛”现象较严重,不同机构间的数据互通共享仍存在不足。此外,约70%的中国罕见肿瘤研究需引用美国数据库数据,如SEER^[27,28]。而在国际数据库访问权限收紧的背景下,国内数据库在罕见肿瘤病种覆盖范围不足、数据结构化程度不高的问题更加凸显。中国虽启动“天枢计划”拟整合资源,但现有数据库的结构化程度仍有待提升,罕见肿瘤病种覆盖有限。

2.2 以儿童小脑髓母细胞瘤生物医学数据库构建为例

中枢神经系统恶性肿瘤是19岁以下人群癌症相关

死亡的首位原因,我国发病率和死亡率均居全球首位^[29]。其中儿童小脑髓母细胞瘤是最常见的中枢神经系统恶性肿瘤之一^[30],但其发病率低于2~3/10万,仍被归类为罕见肿瘤。其复发后五年生存率仅8.2%,但国家和社会对此关注度远远不足。虽然国外单中心患者就诊数量较少,但在美国和加拿大国际知名团队倡导和推动下,已建立了多个高质量的专病数据联盟,不仅包含临床多模态数据(如人口学、病理学、详细治疗方案及随访信息等),还包含了多维组学数据(DNA甲基化芯片数据、转录组测序数据等),在探索相关基因致病机理等方面发挥了关键作用。另外,SEER数据库中也纳入了2675例该疾病信息,但主要集中于流行病学和临床诊疗数据,缺乏大规模多组学生物信息数据;GEO数据库中成规模的研究也仅有2~3项,其他多数是引用此研究中的部分队列,虽然组学数据十分宝贵,但多数样本缺乏完整的治疗及随访记录。

依托首都医科大学附属北京天坛医院及中国国家脑肿瘤注册登记研究平台,研究团队联合国内多中心构建了儿童小脑髓母细胞瘤专病研究队列,涵盖不同分子亚型、详尽诊疗方案等,随访长达20年,其中约70%样本具有多组学数据;并进一步联合美国Texas Children's Hospital领衔的国际多中心队列,纳入近60家医疗中心的795例样本。如图1所示,虽然总体数量较SEER、GEO数据库仍有一定差距,但诊疗信息详细

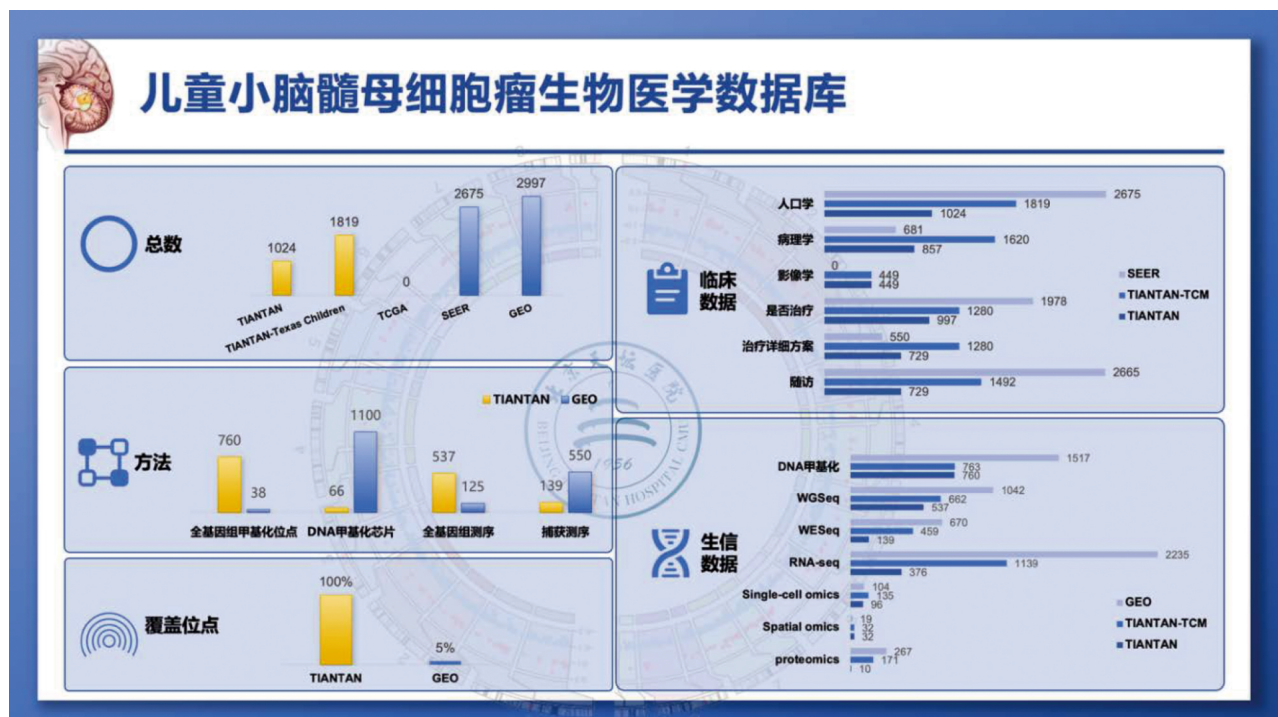


图1 儿童小脑髓母细胞瘤天坛队列与国际数据库比较

Fig.1 Comparison between the Tiantan Pediatric Medulloblastoma Cohort and International Databases

程度、测序深度等质量方面优势明显,并同时配对构建了多种品系实验动物或类器官模型。该数据集的多中心、高质量、高深度特征,可有效支撑下游多维分析与临床转化应用,但其仍存在前瞻性研究队列数据不足的局限性。此外,依托该数据库并结合人工智能技术,已初步构建集数据整合、分子分型、“基因—药效”适配药物筛选和预后风险评估于一体的线上开源平台,供更多研究者和患者使用(图2)。以此罕见肿瘤专病队列生物数据库构建为例,未来将纳入国家罕见病平台,为其他罕见肿瘤数据库的构建提供可复制、可扩展的范式参考。

儿童小脑髓母细胞瘤临床病例分布广泛,不同医疗机构在样本采集、诊断标准、检测流程与数据管理方面存在显著差异。多中心异质性已成为罕见肿瘤数据库建设中不可回避的关键挑战,直接影响数据整合效率与研究结果的可重复性。因此,本项目在组织架构、流程规范与技术手段等方面构建了系统的质量保障机制。在协作组织方面,项目采用“核心牵头—分级协作”的全国多中心模式,由北京天坛医院及国家脑肿瘤注册登记平台负责总体设计与协调,制定统一的数据模板、指标口径和上传规范,设立中央协调办公室负责注册管理、权限分级、技术支持与进度调度。各参与单位通过标准化接口提交临床、组学、影像和随访数据,定期召开线上会议推进项目实施,并通过集中培训与技术答疑提升整

体数据采集质量。

为有效应对多中心样本异质性与数据一致性问题,项目高度重视质量控制体系建设。通过制定统一的病例采集模板、标准化的组学测序流程和数据清洗规则,最大程度降低不同中心在采样、处理、注释环节的系统偏差。所有纳入的组学数据均需通过指定实验平台复核,并采用统一软件流程完成预处理与变异检测,确保数据可比性。临床数据部分设置多级逻辑核查机制,重点检查诊断分型、治疗时间点、复发结局等核心变量的准确性与一致性。此外,依托天坛医院与国际合作团队共同开发的数据平台,在数据上传环节嵌入缺失识别与质量评分系统,实现自动评估与实时反馈,提升整体数据稳定性。

此外,平台建立了周期性反馈机制与多元转化路径。各参与单位每半年接收由协调组发布的数据质量评估报告,内容包括缺失率、逻辑错误率、更新频次等关键指标。评估结果作为后续科研项目参与、数据引用与合作优先级的重要依据,并对表现优异的中心给予专项科研资源支持。研究成果通过联盟年会、联合发表和平台推送等方式持续发布,并逐步应用于人工智能辅助诊断、风险预测与个体化治疗推荐系统的开发,形成“科研—临床”双向闭环。

2.3 实践反思与建议

在多中心数据库构建实践中,团队深刻体会到推进



图2 多模态及多组学数据支撑人工智能辅助的智慧诊疗

Fig.2 Multi-modal and Multi-omics Data Supporting AI-assisted Intelligent Diagnosis and Treatment

数据标准化与样本资源统一面临的现实阻力与操作难点。首先,数据采集模板虽可统一部署,但实际执行中,不同医疗中心在电子病历结构、诊断编码精度及人员理解层面存在较大差异,团队采取“标准模板+人工质控”双轨并行策略,辅以动态培训与集中数据核查,逐步建立起统一采集逻辑。其次,在生物样本库方面,由于各中心样本采集流程和保存条件参差不齐,研究团队制定并推广了适用于全国三级医院的样本采集与冷链转运规范,同时联合病理科与检验科设置前处理质控点,提升样本可用于多组学研究的适配性。针对当前数据库尚缺乏前瞻性队列的问题,项目已在首都医科大学附属北京天坛医院启动以“分层入组、定期随访、样本动态留存”为特征的前瞻性登记计划,并计划逐步拓展至更多区域医疗中心,力图实现从回顾性整合向真实世界长期追踪的转化。

3 应对策略与未来发展方向:交叉融合

3.1 整合优化已有数据,建设自主高质量数据库

我国多个研究中心、三甲医院和区域肿瘤登记体系中已积累大量散在的罕见肿瘤临床及组学数据,但目前存在“信息孤岛”“重复建设”“样本有限”等问题。应加快整合国内分散的罕见肿瘤病例资源,建立国家级统一数据平台,统筹整合高校、科研院所、医疗机构、疾控系统及卫生行政部门的既有数据,构建以病例信息、治疗反应、生存结局和多维组学为核心的综合数据库。应构建真实世界研究体系,利用电子病历、医保数据与随访数据,扩展数据广度^[31]。同时,推动数据“从科研走向产业”,通过政策引导与机制保障,促进科研成果的转化应用与标准落地,推动数据产品应用于临床辅助决策。依托《人类遗传资源管理条例》强化数据主权,推动多家大型三甲医院数据的规范整合。在国际数据封锁壁垒下,应积极发展替代性国际数据资源,可转向欧洲ENA(European Nucleotide Archive)数据库^[32]、日本JGA(Japanese Genotype-phenotype Archive)^[33]等替代资源,同时加速建设国家基因组科学数据中心等自主平台。

3.2 人工智能(AI)技术辅助,提升数据挖掘与诊疗效率

AI可在数据处理、分析及知识挖掘中发挥核心作用,尤其适用于样本量小、异质性强的罕见肿瘤研究。大语言模型应用,结合生成式AI技术,可降低误诊率,缩短诊断周期。整合非结构化临床笔记、基因组数据,多模态数据分析提升数据利用率^[34]。例如,图神经网络与多模态学习可用于跨“影像—组学—病理”数据的联合建模,揭示罕见肿瘤生物学特征;小样本学习与迁移

学习等算法适合解决罕见肿瘤样本数量有限的问题。

3.3 制定标准化数据协议,发展“联邦学习”平台

制定适应中国国情的标准化数据格式和标注协议,如参考临床数据交换标准协会(Clinical Data Interchange Standards Consortium, CDISC)^[35]、美国OMOP(Observational Medical Outcomes Partnership)项目^[36]等。联邦学习(Federated Learning)作为新型协同计算范式,为数据跨机构共享与建模提供了安全可行的解决方案^[37]。应建设区域级或疾病特异性“联邦学习”平台,在不传输原始数据的基础上,联动多中心开展AI建模,确保数据可用不可见,兼顾“数据主权”与“科研开放”双重诉求。另外,推动政府、企业和医疗机构三方协作,构建可扩展的联邦学习生态系统,保障数据合规同时促进科研合作。在联邦学习平台的具体实现路径上,应结合我国现有医疗数据管理实际,探索“中心平台主导+区域节点参与”的分布式建模机制。在安全机制设计方面,可引入联邦平均与加权梯度同步结合的模型更新策略^[38],同时嵌入差分隐私保护与同态加密等隐私计算技术,有效防止原始数据泄露及模型反推攻击^[39,40]。当前,北京、上海、浙江等地已在肿瘤筛查与影像识别等场景中开展区域级联邦学习试点,为罕见肿瘤数据跨机构建模积累了初步经验。

3.4 警惕伦理红线、数据跃进等问题

在数据库快速扩张过程中,应高度警惕数据“跃进式发展”带来的质量隐患,以及伦理与隐私保护风险。数据采集与使用须明确知情同意边界与数据脱敏原则,建立患者数据授权与撤回机制;强化数据库的伦理审查与访问权限控制,避免滥用数据或违反伦理底线^[41]。同时,也应警惕“追求数量、忽视质量”的倾向,避免在规模增长过程中忽视数据的真实性、完整性与可验证性。为确保数据建设质量,应在采集阶段执行统一的结构化模板与编码体系,在整合阶段引入自动逻辑校验、版本追踪与数据溯源机制,在使用阶段建立质量评分体系与可审计的使用记录,形成覆盖数据全生命周期的质量控制闭环^[42]。最后,为改进当前伦理与数据共享机制,应推动建立“快速伦理通道+集中评审授权”机制,在不涉及干预性研究前提下,探索科研豁免路径与多中心伦理互认模式。建议构建覆盖国家—区域—机构三级的数据授权治理平台,配套区块链式数据访问日志记录系统,实现“可溯源、可问责、可审计”的共享监管结构。

4 总结与展望

罕见疾病,尤其儿童中枢神经系统肿瘤,面临致病

机制复杂、组织学异质性高、诊治手段有限及预后差等严峻挑战,高度依赖跨学科、跨机构的数据协作。高质量生物医学数据集是推动罕见疾病研究的关键基础,可支撑分子分型、靶点识别及药物开发等。然而,罕见肿瘤数据集仍存在病种覆盖不足、标准制定杂乱、共享机制缺失等问题,而且近年国际数据共享环境恶化,更加阻碍数据共享推进。本文以儿童小脑髓母细胞瘤为例,探讨罕见疾病数据库的构建策略。依托多中心专病队列,整合临床、多组学数据及AI分析平台,为研究提供高质量资源。未来需通过优化数据整合、强化AI技术应用、制定标准化协议及发展联邦学习平台,破解数据孤岛问题,同时需平衡数据共享与伦理隐私风险,推动我国自主罕见肿瘤数据库体系建设。

参 考 文 献

- [1] Boycott KM, Giugliani R, Commissioners of the RDI-Lancet Commission on Rare Diseases. The RDI-lancet commission on rare diseases: improving visibility to address health-care disparities for 400 million people. *Lancet*, 2025, 405(10479): 605—607.
- [2] Elmadani M, Klara S, Mustafa M, et al. Global burden of rare cancers: insights from GLOBOCAN 2022 estimates. *Cancers*, 2025, 17(10): 1721.
- [3] Cohen AR. Brain tumors in children. *The New England Journal of Medicine*, 2022, 386(20): 1922—1931.
- [4] Cianflone A, Savoia F, Parasole R, et al. Pediatric biobanks to enhance clinical and translational research for children. *European Journal of Pediatrics*, 2023, 182(4): 1459—1468.
- [5] CNCB-NGDC Members and Partners. Database resources of the national genomics data center, China national center for bioinformatics in 2023. *Nucleic Acids Research*, 2023, 51(D1): D18—D28.
- [6] Greene D, Pirri D, Frudd K, et al. Genetic association analysis of 77, 539 genomes reveals rare disease etiologies. *Nature Medicine*, 2023, 29(3): 679—688.
- [7] Tryka KA, Hao LN, Sturcke A, et al. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Research*, 2014, 42(Database Issue): D975—D979.
- [8] Wang ZN, Jensen MA, Zenklusen JC. A practical guide to the cancer genome atlas (TCGA). *Methods in Molecular Biology*, 2016, 1418: 111—141.
- [9] Schatz MC, Philippakis AA, Afgan E, et al. Inverting the model of genomics data sharing with the NHGRI genomic data science analysis, visualization, and informatics lab-space. *Cell Genomics*, 2022, 2(1): 100085.
- [10] Guo J, Liu P, Chen LM, et al. National Rare Diseases Registry System (NRDRS): China's first nation-wide rare diseases demographic analyses. *Orphanet Journal of Rare Diseases*, 2021, 16(1): 515.
- [11] Subbiah V, Othus M, Palma J, et al. Designing clinical trials for patients with rare cancers: connecting the zebras. *American Society of Clinical Oncology Educational Book American Society of Clinical Oncology Annual Meeting*, 2025, 45(3): e100051.
- [12] Gatta G, van der Zwan JM, Casali PG, et al. Rare cancers are not so rare: the rare cancer burden in Europe. *European Journal of Cancer*, 2011, 47(17): 2493—2511.
- [13] Siegel RL, Kratzer TB, Giaquinto AN, et al. Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, 2025, 75(1): 10—45.
- [14] 岳和欣, 湛永乐, 边峰, 等. 临床队列研究的数据标准与共享. *中华流行病学杂志*, 2021, 42(7): 1299—1305.
- Yue HX, Zhan YL, Bian F, et al. Data standard and data sharing in clinical cohort studies. *Chinese Journal of Epidemiology*, 2021, 42(7): 1299—1305. (in Chinese)
- [15] 郭健, 刘鹏, 荆志成, 等. 中国国家罕见病注册系统建设及应用. *罕见病研究*, 2022, 1(1): 7—12.
- Guo J, Liu P, Jing ZC, et al. Construction and application of national rare diseases registry system of China. *Journal of Rare Diseases*, 2022, 1(1): 7—12. (in Chinese)
- [16] Raab R, Küderle A, Zakreuskaya A, et al. Federated electronic health records for the European health data space. *The Lancet Digital Health*, 2023, 5(11): e840—e847.
- [17] Pillai RK, Jayasree K. Rare cancers: challenges & issues. *Indian Journal of Medical Research*, 2017, 145(1): 17—27.
- [18] Tiwari A, Mishra S, Kuo TR. Current AI technologies in cancer diagnostics and treatment. *Molecular Cancer*, 2025, 24(1): 159.
- [19] Nishat SMH, Shahid Tanweer A, Alshamsi B, et al. Artificial intelligence: a new frontier in rare disease early diagnosis. *Cureus*, 2025, 17(2): e79487.
- [20] Yang HX, Xiong J, Zhao WL. Advancements in artificial intelligence for the precise diagnosis and treatment of hematological malignancies. *Zhonghua Xueyexue Zazhi*, 2025, 46(2): 186—192.
- [21] Christyani G, Carswell M, Qin SS, et al. An overview of advances in rare cancer diagnosis and treatment. *International Journal of Molecular Sciences*, 2024, 25(2): 1201.
- [22] Lai JK, Liao XW, Yao C, et al. Existing barriers and recommendations of real-world data standardisation for clinical research in China: a qualitative study. *BMJ Open*, 2022, 12(8): e059029.
- [23] Bradford D, Reilly KM, Widemann BC, et al. Developing therapies for rare tumors: opportunities, challenges and progress. *Expert Opinion on Orphan Drugs*, 2016, 4(1): 93—103.
- [24] Casali PG, Bruzzi P, Bogaerts J, et al. Rare Cancers Europe (RCE) methodological recommendations for clinical studies in rare cancers: a European consensus position paper. *Annals of Oncology*, 2015, 26(2): 300—306.
- [25] Mansouri A, Zadeh G. NCI-CONNECT rare CNS tumor initiative: current progress and anticipated future impact. *Neuro-Oncology Advances*, 2020, 2(1): vdaa022.
- [26] Xue JX, Lyu QY. Challenges and opportunities in rare cancer research in China. *Science China Life Sciences*, 2024, 67(2): 274—285.
- [27] Huang CD, Huang JG, He Y, et al. Competing-risks model for predicting the prognosis of patients with angiosarcoma based on the SEER database of 3905 cases. *Holistic Integrative Oncology*, 2024, 3(1): 13.
- [28] Xia LJ, Lai JQ, Liu XX, et al. Epidemiological and survival outcomes

- of neuroendocrine carcinoma of the breast;a SEER data analysis. *Translational Cancer Research*,2023,12(8):1951—1962.
- [29] Patel AP,Fisher JL,Nichols E,et al. Global,regional,and national burden of brain and other CNS cancer,1990–2016;a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*,2019,18(4):376—393.
- [30] Aref D,Croul S. Medulloblastoma;recurrence and metastasis. *CNS Oncology*,2013,2(4):377—385.
- [31] Mendez KM,Reinke SN,Kelly RS,et al. A roadmap to precision medicine through post-genomic electronic medical records. *Nature Communications*,2025,16:1700.
- [32] Yuan D,Ahamed A,Burgin J,et al. The European nucleotide archive in 2023. *Nucleic Acids Research*,2024,52(D1):D92—D97.
- [33] Kodama Y,Ara T,Fukuda A,et al. DDBJ update in 2024;the DDBJ Group Cloud service for sharing pre-publication data. *Nucleic Acids Research*,2025,53(D1):D45—D48.
- [34] Waqas A,Tripathi A,Ramachandran RP,et al. Multimodal data integration for oncology in the era of deep neural networks;a review. *Frontiers in Artificial Intelligence*,2024,7:1408843.
- [35] Facile R,Muhlbradt EE,Gong MC,et al. Use of clinical data interchange standards consortium (CDISC) standards for real-world data;expert perspectives from a qualitative Delphi survey. *JMIR Medical Informatics*,2022,10(1):e30363.
- [36] Williams N. Building the observational medical outcomes partnership's T-MSIS Analytic File common data model. *Informatics in Medicine Unlocked*,2023,39:101259.
- [37] Choudhury A,Volmer L,Martin F,et al. Advancing privacy-preserving health care analytics and implementation of the personal health train: federated deep learning study. *Jmir Ai*,2025,4:e60847.
- [38] Sun T,Li DS,Wang B. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,2023,45(4):4289—4301.
- [39] Ficek J,Wang W,Chen HN,et al. Differential privacy in health research;a scoping review. *Journal of the American Medical Informatics Association*,2021,28(10):2269—2276.
- [40] Yang WC,Wang S,Cui H,et al. A review of homomorphic encryption for privacy-preserving biometrics. *Sensors*,2023,23(7):3566.
- [41] Ige AB,Chukwurah N,Idemudia C,et al. Ethical considerations in data governance:Balancing privacy,security,and transparency in data management. *International Journal of Engineering Research and Development* 2024. 20:462—468.
- [42] Said A,Rasel FM. Data governance strategies for data quality and integrity in clinical trials. 2023. <https://www.researchgate.net/publication/373214405>.

Establishing a High-quality Biomedical Resource for Rare Diseases: Insights from Pediatric Medulloblastoma

Hailong Liu^{1, 2, 3} Chengyu Ma⁴ Tao Jiang^{1, 2, 5} Yiying Li^{1, 5, 6} Rui Chen⁴ Yu Tian^{4*}

1. *Beijing Tiantan Hospital Capital Medical University, Beijing 100070, China*

2. *National Center for Neurological Disorders, Beijing 100070, China*

3. *Chinese Institute for Medical Research, Beijing 100069, China*

4. *School of Public Health Capital Medical University, Beijing 100069, China*

5. *Beijing Neurosurgical Institute, Beijing 100070, China*

6. *Beijing Engineering Research Center of Targeted Drugs and Cell Therapy for CNS Tumors, Beijing 102600, China*

Abstract Rare cancers, despite their low individual incidence, encompass a wide array of subtypes and collectively account for more than 20% of all malignancies. Progress in this field has been constrained by limited sample availability and fragmented research resources, making precision diagnosis and targeted therapy heavily reliant on high-quality, openly accessible databases. While recent advances in genomic technologies have facilitated progress in selecting rare cancer types, the persistence of data silos and insufficient data integration remains a critical barrier to the development of precision oncology. Taking pediatric medulloblastoma as the insight, we examine the current state of biomedical data infrastructure for rare cancers, assess the impact of data fragmentation on research productivity, and outline strategic recommendations for building robust, high-resolution databases to accelerate discovery and improve patient outcomes.

Keywords rare cancer; tumor database; precision medicine; artificial intelligence; medulloblastoma

刘海龙 首都医科大学附属北京天坛医院副主任医师,副教授。主要从事神经系统肿瘤的综合诊疗与转化应用研究。发表学术论文50余篇,主持国家自然科学基金项目、国家科技重大专项和北京市杰出青年科学基金等科研项目10余项,授权国家发明专利2项。

田雨 首都医科大学公共卫生学院副教授。主要从事慢性病流行病学及健康医疗大数据相关研究。主持或参与国家自然科学基金项目及多项国家级、省部级课题。在*British Medical Journal*、*Gastroenterology*、*Journal of the National Cancer Institute*等期刊发表学术论文30余篇,授权国家发明专利2项。

(责任编辑 陈鹤 张强)

* Corresponding Author, Email: yutian@ccmu.edu.cn