· 述评 ·

胃肠肿瘤医疗大数据库的机遇与挑战

季加孚[1]　何琦非[1]　王晓云[2]　于文博[2]　苗儒林[1]　应项吉[1]　卢新璞[1]
[1]北京大学肿瘤医院暨北京市肿瘤防治研究所胃肠肿瘤中心 恶性肿瘤发病机制及转化研究教育部重点实验室　100142；[2] 医渡云(北京)技术有限公司　100101
通信作者：季加孚，Email：jijiafu@hsc.pku.edu.cn

【摘要】 随着信息化技术的发展，大数据时代的到来，国家出台了多部政策纲要指导包括医疗在内的多个行业大数据应用和发展。笔者介绍了医疗大数据发展的时代背景和重要意义，回顾总结了国外大数据平台的特点，对医疗大数据平台的管理和应用进行阐述，对胃肠肿瘤乃至整个医疗行业的大数据发展进行展望和憧憬。

【关键词】 胃肿瘤；　肠肿瘤；　医疗大数据；　精准医疗；　数据库

基金项目：北京市医院管理局培育计划(PX2018043)

DOI：10.3760/cma.j.issn.1673-9752.2019.03.001

Opportunities and challenges of medical big database of gastrointestinal tumor

Ji Jiafu[1], *He Qifei*[1], *Wang Xiaoyun*[2], *Yu Wenbo*[2], *Miao Rulin*[1], *Ying Xiangji*[1], *Lu Xinpu*[1]
[1]*Gastrointestinal Cancer Center, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing 100142, China*; [2]*Yiducloud (Beijing) Technology Co., Ltd. Beijing 100101, China*
Corresponding author: Ji Jiafu, Email: jijiafu@hsc.pku.edu.cn

【**Abstract**】 With the development of information technology and the arrival of the era of big data, our country has introduced a number of policies and regulations to guide the application and development of big data in many industries including health care. This article introduced the background and significance of the development of medical big data, reviewed the characteristics of foreign big data platforms, discussed the management and application of medical big data platform, and anticipated the future development of big data for gastrointestinal cancer and even the entire medical industry.

【**Key words**】 Gastric neoplasms; Intestinal neoplasms; Medical big data; Precision medicine; Database

Fund program: Development Program of Beijing Municipal Administration of Hospitals (PX2018043)

DOI：10.3760/cma.j.issn.1673-9752.2019.03.001

随着互联网技术、物联网、云计算等信息化的发展，“大数据”时代已经到来。2015 年，十八届五中全会首次提出“国家大数据战略”[1]。2016 年 7 月中共中央办公厅、国务院办公厅印发了《国家信息化发展战略纲要》[2]。2016 年国务院先后印发了《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》(以下简称《意见》)和《“健康中国 2030”规划纲要》，提出我国要规范和推动健康医疗大数据融合共享、开放应用，全面深化健康医疗大数据在行业治理、临床和科研、公共卫生等领域的应用，培育健康医疗大数据应用新业态，医疗行业正式步入大数据时代[3-4]。

1　医疗大数据平台的重要性

医疗大数据是指在医疗行业中产生的数据，具有大数据的 4V 属性，即大规模(volume)、多样性(variety)、产生和变化速度快(velocity)和价值密度低(value)。其来源于临床医疗或检验检查数据、制药企业、医疗和医保费用管理、健康管理或社交网络等，具有“分散化、碎片化、透明度低、语义复杂”等特征[5]。如何解决医疗数据的完整性、准确性、一致性问题，把这些纷繁的单体原始数据变成互联互通的“可计算、可使用”数据[6]。这是传统医疗数据平台面临的挑战，也是精准医疗时代对数据标准的要求[7-8]。在肿瘤领域，越来越多的高水平研究建立在上万病例数据的分析基础上，从肿瘤发生机制的研究到临床诊断与治疗、预防和监测，都涉及数据收集、管理和分析等[9]。因此，高质量的临床大数据是未来精准临床决策与高水平临床研究的首要基础条件。构建全面完善的肿瘤医疗大数据平台，以大数据智能化驱动基于真实世界数据的医学研究，加强患者健康状况等重要数据精准统计和预测评价，建设好人民满意的医疗卫生事业，为打造健康中国、全面建成小康社会和实现中华民族伟大复兴的中国梦提供有力支撑[10]。

2　国外经验

肿瘤数据的收集、分析一直是各国肿瘤研究乃

至促进国民健康的重点，高质量、大规模大样本的肿瘤数据库建立成为各国肿瘤研究的基础。目前，国外已有比较成熟的大型肿瘤数据库。他山之石，可以攻玉。这些数据库的发展、演变、运行可以为我国大数据平台的建立提供经验。

美国监测、流行病学和结果（Surveillance, Epidemiology, and End Results, SEER）数据库是 1973 年由美国国家癌症研究所建立，用于监测肿瘤发病情况、分期、治疗以及预后信息的数据库[11-12]。美国国家癌症数据库（National Cancer Database, NCDB）始于 1988 年，由美国外科医师学会与美国癌症协会建立，是以医院为基础的癌症登记系统[13-14]。日本癌症登记数据库是由日本国家癌症中心负责建立，分为基于人群和基于医院的癌症登记，由《日本癌症肿瘤登记促进法案》和《癌症登记法》强制保障其癌症登记的实施和落实[15-16]。日本的国家临床数据库（National Clinical Database, NCD）则是一个全国性的外科手术数据录入电子系统，于 2010 年在日本外科协会和日本胃肠外科协会等的支持下启动，同时 NCD 还与外科协会的认证系统相关联，用于评价协会认证的外科医师[17-18]。见表 1。

目前国际大型肿瘤数据库大致分为两类：基于人群的数据库和基于医院系统的数据库。美国 SEER 数据库和日本癌症登记数据库属于前者，而美国的 NCDB 和日本的 NCD 属于后者。基于医院的数据库，更符合临床需求，可为患者诊断与治疗提供更多针对性的信息。但是这类数据库存在就诊偏倚，并不能很好地反映人口学分层的特点，例如 NCDB 和 SEER 数据库在某些癌种的种族、年龄分布上存在差异[19]。基于人群的数据库流行病学意义更加明确，能为国家战略制定提供更多依据。通常两种形式数据库间的相互融合、数据共享能起到 1+1>2 的作用。例如，日本 NCD 主要收集详尽的围术期数据，而随访数据的积累一定程度上需依靠癌症登记数据库完成。由于《癌症登记法》的强制性和广泛覆盖，肿瘤登记处会收集肿瘤患者的预后信息。这些信息会由登记处返回到提供信息的医院，NCD 即可通过医院获取肿瘤相关预后信息[16]。此外，NCD 还和日本诊断程序组合数据库（Diagnosis Procedure Combination Database, DPCD）等医疗保险数据库互通，有利于开展卫生经济学相关研究[20]。与之类似，美国 SEER 数据库与医疗保险合作，形成了 SEER-Medicare 数据库[12]。

需要说明的是，日本一些学术团体和研究机构建立的肿瘤专病数据库在 NCD 出现之前就已经建立并独立运行了很长时间。1952 年日本妇科肿瘤数据库是第一个癌症专病数据库，此后胃癌（1963 年）、食管癌（1965 年）、肝癌（1965 年）等肿瘤先后建成了相应的专病数据库，目前仍有相关的随访数据发表[21]。但是，这些专病数据库存在覆盖率低、管理不稳定、标准化程度低等问题，基于专病数据库的建立和运营经验，日本 NCD 等肿瘤综合数据库逐步建立。鉴于 NCD 出色的组织和管理水平，胰腺癌、乳腺癌和肝癌等专病数据库逐渐并入了 NCD[16]。

总体而言，由学术团体专病数据库，到全国性数据库，再到全国不同数据库间的数据共享整合是国外大数据发展的脉络和趋势。如何促进数据间的开放，避免数据孤岛的形成，避免信息资源的浪费，是国内研究者亟需解决的问题。

表 1　美国和日本肿瘤数据库情况

数据库名称	成立机构	成立时间	数据库规模	类别	数据内容
美国监测、流行病学和结果数据库	美国国家癌症研究所	1973 年	910 万条记录，覆盖了 28% 的美国人口	基于人群的癌症登记系统	监测肿瘤发病情况、分期、治疗以及预后信息
美国国家癌症数据库	美国外科医师学会与美国癌症协会	1988 年	截至 2016 年，累计达 3 400 万条记录，覆盖美国 30% 的医院	基于医院的癌症登记系统	人口特征、肿瘤特征、治疗及治疗并发症相关信息
日本癌症登记数据库	日本国家癌症中心	2013 年	2015 年报告中，745 个登记处覆盖了日本所有的县，包涵了 80% 以上的肿瘤患者	基于人群和基于医院的登记系统同时存在	人口特征、疾病相关特征、随访信息
日本国家临床数据库	日本外科协会和日本胃肠外科协会	2010 年	涵盖了日本≥95% 的常规手术，截至 2016 年 12 月，共有 5 000 余家机构注册，9 100 000 例患者完成登记	基于医院的外科手术数据录入电子系统，与外科医师资质认证相关	手术相关详细信息，在胃肠外科方面包括：食管切除术、远端胃切除术、全胃切除术、右半结肠切除术等 8 个路径，涵盖了 115 种胃肠手术方式；2016 年开始收集并发症数据

3　国内数据平台建立的挑战与应对策略

我国作为胃肠肿瘤大国,患者规模位于全世界之首,与之相对应的大数据资源也应冠于全世界。整合资源建立大数据平台,能提高对疾病风险因素的分析、预测、防范能力,全面提升我国肿瘤患者,乃至整体国民的健康水平。然而当前情况却不尽如人意,胃肠肿瘤乃至大多数病种的数据库都存在着“小”“差”“乱”的情况。“小”指的是很多数据库规模小、病例数少,数据条目结构少。“差”指的是数据质量差,尤其是我国由于人口流动性大,随访数据难以收集,存在数据丢失的情况。“乱”是指数据一致性差,数据库融合共享可能性低。

3.1　机遇与优势

诚然我们的医疗数据库存在着各种各样的不足,面临着各种各样的困难,高质量数据库的建立与发达国家存在一定差距。但同时我们也处在最好的时代,掌握着迎头赶超发达国家数据平台的机遇和优势。

政策导向方面,2015 年国家大数据战略出台了《意见》等相关指导纲要,要求大数据建设要“坚持规范有序、安全可控;坚持开放融合、共建共享”。为顺应时代发展趋势,由北京大学肿瘤医院牵头成立“中国胃肠肿瘤外科联盟”(简称胃肠联盟),启动了胃癌专病标准化数据库的建立,针对胃癌、肠癌的基本临床数据及并发症相关数据进行收集[22]。通过一致性可供推广的胃癌临床数据库标准,实现了跨部门、跨医院、跨系统的互联互通和信息共享平台,推动了我国胃肠肿瘤外科的合作、交流与进步。

时代发展方面,数据库的发展,离不开技术的进步。2016 年美国 NCDB 线上共享文件(Participant User File, PUF)的普及,使得基于 NCDB 3 年发表的研究数量几乎与 SEER 数据库和 SEER-Medicare 数据库相当[23]。欧美国家、日本和韩国高质量数据平台建立和数据收集得力于互利网和个人计算机的发展,那么我们有理由相信随着我国移动互联网技术、通信技术的发展,我们能在未来医疗数据平台的建设中占有更多主动权。目前我国通信基础设施的进步已经成为了我们探索“互联网+健康医疗”模式的先决条件。借助各类社交软件及平台,肿瘤患者的随访跟踪可开展线上线下结合的医疗服务新模式,通过民间创新力量的合作介入,利用微信等常规通讯软件进行医疗数据存储、清洗等。利用移动通信的发展,充分发挥患者能动性,及时获取患者状态,可对我们的临床研究和患者管理起到事半功倍的作用。

3.2　挑战与应对

医疗大数据平台的建立和运营,存在医疗和技术两方面的困难。优质的平台需要医疗行业与技术力量的深度融合。标准化、平台更新、质量控制等一直是困扰医疗数据平台运营的难题。而数据的存储清洗、结构化处理、分析挖掘、安全隐私保护等技术也需要重点攻关。

3.2.1　医疗数据库建立的困难:数据平台建立的挑战来自于医学的快速发展。肿瘤分期、分型不断更新,个体化治疗需要的肿瘤分子生物学信息、各层面的组学信息也日益增多。数据库收录信息也要与时俱进,但数据库结构的变化势必会影响已收集信息的准确性和完整性。因此,数据库的使用者应该熟悉数据库每一个版本的更新演变内容,清楚每项条目的准确意义;此外,数据库需要详尽收集原始数据,例如 NCDB 的协同分期系统要求分期方面收集足够信息,以满足不同分期系统的要求[14]。依托于原始数据的平台,对于新的肿瘤分期、分型系统可以进行快速验证[24]。作为数据的原始资料,临床医师要重视病历的质量,详细的病历记录可以保证数据回顾的准确性。美国奥巴马政府 2009 年通过的《经济与临床医疗信息技术法案》要求所有数据不论好坏都要存储。这与我国胃肠肿瘤外科术后并发症登记和质量控制的要求一致[25]。

3.2.2　技术难题与解决:当前,很多医院已完成电子病历系统、实验室检查系统、影像系统等建设工作,产生了庞大且繁杂的医疗数据,但医院对现有数据的利用度极低。由于各医院独立运营,数据库规则不尽相同,医院间相互协作业务互联互通存在一定困难。与民间资本的深度合作可以在一定程度上缓解这一难题。结合行业标准,搭建非结构化数据转换的标准与规范体系,利用知识图谱和自然语言处理等技术,让经验丰富的医学专家与数据科学家、IT 互联网专业人士进行全方位的沟通与合作,创建“医学数据智能平台”,对大规模多源异构医疗数据进行抽取、转换、标准化、结构化、归一,形成基于诊断与治疗全流程的临床病患管理医学数据。参考国际通用的健康保险携带和责任(HIPAA)法案对患者数据进行脱敏,保证患者数据隐私;采用加密强度较高的算法,确保数据存储与传输的安全问题;参照国家信息安全等级保护,引进吸收国外医疗行业先进数据安全管理理念,实现传统网络安全与数据安全的融合。

4　结语

新一代测序技术的发展使得精准治疗成为可能，大量的生物信息学信息、组学信息，甚至环境、生活方式的信息都能获取并加以分析，能为临床医师的研究和临床治疗提供更多信息。日本有针对消化道肿瘤开展的全国性 GI-SCREEN-Japan 基因组筛查项目，欧洲癌症治疗研究组织（EORTC）开展了针对结直肠癌的 SPECTAcolor 精准治疗项目，美国国家癌症研究所（NCI）发起的 MATCH 项目[26]。通过全面整合数据，建立以患者为中心的数据网络，提供疾病分类、诊断、治疗发展、临床决策、药物研发等所需信息；基于数据库的深度处理和分析，建立真实世界疾病领域模型，构建精准医疗知识网络，全方位提升这些数据的价值，使医疗大数据高效助力精准医疗服务、医疗管理、政府公共决策、创新新药研发等，引领大健康产业创新，实现数据智能绿色医疗的新生态。

笔者也期待在不久的将来，通过政产学结合的方式，促成临床、科研和公共卫生大数据资源的共享和开放，全面提升医学科研、医疗服务的质量和效率。笔者相信在未来，随着我国居民健康、社会保障等数据的应用集成，能更方便地获取覆盖居民全生命周期的健康信息，通过和我们已有数据库的共享，建设符合我国国情的大数据平台，逐步形成具有国际影响力的大型胃肠肿瘤数据库。

利益冲突　所有作者均声明不存在利益冲突

参　考　文　献

[1] 中共中央关于制定国民经济和社会发展第十三个五年规划的建议[N].人民日报，2015-11-04(001).

[2] 中共中央办公厅、国务院办公厅印发《国家信息化发展战略纲要》[EB/OL].(2016-07-27)[2019-03-01]. http://www.gov.cn/xinwen/2016-07/27/content_5095297.htm.

[3] 国务院办公厅印发《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].(2016-06-24)[2019-03-01]. http://www.gov.cn/xinwen/2016-06/24/content_5085211.htm.

[4] 曾钊，刘娟.中共中央、国务院印发《"健康中国 2030"规划纲要》[J].中华人民共和国国务院公报，2016，(32)：5-20.

[5] Austin C, Kusumoto F. The application of Big Data in medicine: current implications and future directions[J]. J Interv Card Electrophysiol, 2016, 47(1): 51-59. DOI: 10.1007/s10840-016-0104-y.

[6] Botsis T, Hartvigsen G, Chen F, et al. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities[J]. Summit Transl Bioinform, 2010, 2010: 1-5.

[7] Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data[J]. EGEMS (Wash DC), 2016, 4(1): 1244. DOI: 10.13063/2327-9214.1244.

[8] McCowan C, Thomson E, Szmigielski CA, et al. Using Electronic Health Records to Support Clinical Trials: A Report on Stakeholder Engagement for EHR4CR[J]. Biomed Res Int, 2015, 2015: 707891. DOI: 10.1155/2015/707891.

[9] 周殷杰，向明飞，李涛.医疗大数据在恶性肿瘤诊治中的应用[J].国际肿瘤学杂志，2016，43，(1)：75-78. DOI: 10.3760/cma.j.issn.1673-422X.2016.01.021.

[10] Shah A, Stewart AK, Kolacevski A, et al. Building a Rapid Learning Health Care System for Oncology: Why CancerLinQ Collects Identifiable Health Information to Achieve Its Vision[J]. J Clin Oncol, 2016, 34(7): 756-763. DOI: 10.1200/JCO.2015.65.0598.

[11] National Cancer Institute: surveillance, epidemiology, and end results program[DB/OL].[2019-03-01]. https://seer.cancer.gov.

[12] Park HS, Lloyd S, Decker RH, et al. Overview of the Surveillance, Epidemiology, and End Results Database: Evolution, Data Variables, and Quality Assurance[J]. Curr Probl Cancer, 2012, 36(4): 183-190. DOI: 10.1016/j.currproblcancer.2012.03.007.

[13] National Cancer Database[DB/OL].[2019-03-01]. https://www.facs.org/quality-programs/cancer/ncdb.

[14] Boffa DJ, Rosen JE, Mallin K, et al. Using the National Cancer Database for Outcomes Research: A Review[J]. JAMA Oncol, 2017, 3(12): 1722-1728. DOI: 10.1001/jamaoncol.2016.6905.

[15] National Cancer Center Japan[DB/OL].[2019-03-01]. https://www.ncc.go.jp/en/cis/divisions/stat/index.html.

[16] Anazawa T, Miyata H, Gotoh M. Cancer registries in Japan: National Clinical Database and site-specific cancer registries[J]. Int J Clin Oncol, 2015, 20(1): 5-10. DOI: 10.1007/s10147-014-0757-4.

[17] National Clinical Database[DB/OL].[2019-03-01]. http://www.ncdorjp.

[18] Seto Y, Kakeji Y, Miyata H, et al. National Clinical Database (NCD) in Japan for gastroenterological surgery: Brief introduction[J]. Ann Gastroenterol Surg, 2017, 1(2): 80-81. DOI: 10.1002/ags3.12026.

[19] Lerro CC, Robbins AS, Phillips JL, et al. Comparison of Cases Captured in the National Cancer Data Base with Those in Population-based Central Cancer Registries[J]. Ann Surg Oncol, 2013, 20(6): 1759-1765. DOI: 10.1245/s10434-013-2901-1.

[20] Yasunaga H, Hashimoto H, Horiguchi H, et al. Variation in cancer surgical outcomes associated with physician and nurse staffing: a retrospective observational study using the Japanese Diagnosis Procedure Combination Database[J]. BMC Health Serv Res, 2012, 12: 129. DOI: 10.1186/1472-6963-12-129.

[21] Matsuno S, Egawa S, Fukuyama S, et al. Pancreatic Cancer Registry in Japan: 20 years of experience[J]. Pancreas, 2004, 28(3): 219-230.

[22] "中国胃肠肿瘤外科联盟"成立并发布初步数据[J].中国实用外科杂志，2016，36(10)：1077.

[23] Institute NC. SEER-Medicare publications by journal & year[DB/OL].(2015-03-02)[2019-03-01]. https://health.caredelivery.cancer.gov/seermedicare/overview/pubs_jour_year.php.

[24] Li Z, Wang Y, Fei S, et al. ypTNM staging after neoadjuvant chemotherapy in the Chinese gastric cancer population: an evaluation on the prognostic value of the AJCC eighth edition cancer staging system[J]. Gastric Cancer, 2018, 21(6): 977-987. DOI: 10.1007/s10120-018-0830-1.

[25] 中国胃肠肿瘤外科联盟，中国抗癌协会胃癌专业委员会.中国胃肠肿瘤外科术后并发症诊断登记规范专家共识(2018 版)[J].中国实用外科杂志，2018，38(6)：589-595. DOI: 10.19538/j.cjps.issn1005-2208.2018.06.01.

[26] Bando H. The current status and problems confronted in delivering precision medicine in Japan and Europe[J]. Curr Probl Cancer, 2017, 41(3): 166-175. DOI: 10.1016/j.currproblcancer.2017.02.003.

（收稿日期：2019-03-05）

本文引用格式

季加孚，何琦非，王晓云，等.胃肠肿瘤医疗大数据库的机遇与挑战[J].中华消化外科杂志，2019，18(3)：199-202. DOI: 10.3760/cma.j.issn.1673-9752.2019.03.001.

Ji Jiafu, He Qifei, Wang Xiaoyun, et al. Opportunities and challenges of medical big database of gastrointestinal tumor[J]. Chin J Dig Surg, 2019, 18(3): 199-202. DOI: 10.3760/cma.j.issn.1673-9752.2019.03.001.