doi:10.3971/j.issn.1000-8578.2024.23.1224

• 特约来稿 •

**陈凯** 医学博士，主任医师，硕士生导师，现任中山大学孙逸仙纪念医院乳腺肿瘤中心乳腺外科副主任，中山大学孙逸仙纪念医院人工智能研究室副主任。美国Johns Hopkins Hospital访问学者，广东省医学会外科学分会青年委员会副主任委员，广东省医疗行业协会乳腺专科管理分会副主任委员，广东省医疗行业协会乳腺病整形修复管理分会副主任委员，国家卫生健康委医院管理研究所乳腺癌单病种诊疗能力提升项目专家委员会副秘书长，广东省医学会人工智能分会常委，广东省抗癌协会乳腺癌专业委员会委员，广东省抗癌协会肿瘤标志专业委员会委员。

# 下一代通用型临床科研专病数据库的建设理念框架思考

彭敏[1,2]，陈倩茹[1,3]，陈凯[1,3,4]

**Conceptual Framework of A Next-generation General-purpose Clinical Scientific Disease Database**

PENG Min[1,2], CHEN Qianru[1,3], CHEN Kai[1,3,4]

*1. Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China; 2. Disease Registry Department, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China; 3. Breast Tumor Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China; 4. Artificial Intelligence Lab, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China*

*Corresponding Author: CHEN Kai, E-mail: chenkai23@mail.sysu.edu.cn*

**Abstract:** Empowered by the rapid development of artificial intelligence technology and information security technology, basic medicine, translational medicine, and clinical medicine have entered a new era. In this process, the relevant data such as text, genetics, images, and videos have increased at an astonishing rate. Traditional specialized disease databases cannot adapt to the multimodality of data and the multicentralization of data sources in the present/future. Therefore, establishing a next-generation multimodal and multicenter collaborative specialized disease database that is suitable for secure data sharing and ethical in clinical research is essential. This paper expounds the concept and basic framework of developing a next-generation clinical research disease database with universal use, emphasizing the establishment of a multimodal data integration mechanism, a multicenter collaborative system supported by privacy-preserving computing, highly intelligent software and hardware entities, and an integrated operational management mode. It also introduces the application of this concept in the construction of a breast cancer three-dimensional life database, a project of China Breast Cancer Innovation Alliance.

**Key words:** Specialized disease data; Data management; Clinical research

**收稿日期**：2023-11-23；**修回日期**：2024-01-18

**基金项目**：国家自然科学基金（82271650）；广东省发改委大数据工程实验室项目（4250022021）；广州市科技计划项目（202102010221，20212200003）；国家卫生健康委员会医院管理研究所课题（RXDBZ-2022-04）；中山大学孙逸仙纪念医院逸仙管理项目（GL2307）

**作者单位**：1. 510120 广州，中山大学孙逸仙纪念医院 广东省恶性肿瘤表观遗传与基因调控重点实验室；2. 510120 广州，中山大学孙逸仙纪念医院随访中心；3. 510120 广州，中山大学孙逸仙纪念医院乳腺肿瘤中心；4. 510120 广州，中山大学孙逸仙纪念医院人工智能研究室

**通信作者**：陈凯（1985-），男，博士，主任医师，主要从事乳腺肿瘤、人工智能的研究，E-mail: chenkai23@mail.sysu.edu.cn，ORCID: 0000-0002-7052-4681

**作者简介**：彭敏（1985-），女，硕士，主管技师，主要从事肿瘤登记、数据管理工作，ORCID: 0000-0002-8755-7121

database; Database construction; Breast cancer
**Funding:** National Natural Science Foundation of China (No. 82271650); Guangdong Development and Reform Commission (No. 4250022021); Guangzhou Science Technology and Innovation Commission (Nos. 202102010221, 20212200003); National Institute of Hospital Administration, National Health Commission of the People's Republic of China (No. RXDBZ-2022-04); Yat-sen Management Program of Sun Yat-sen Memorial Hospital, Sun Yat-sen University (No. GL2307)
**Competing interests:** The authors declare that they have no competing interests.

**摘　要：**基础医学、转化医学和临床医学在快速发展的人工智能技术与信息安全技术的赋能下，进入了一个全新的时代。在此过程中，相关的文本、基因、图像及视频等数据以惊人的速度增长。传统的专病数据库无法适应当下以及未来数据多模态化和数据来源多中心化的特点。建设下一代多模态多中心协作的、适合安全数据共享且符合伦理规范的临床科研专病数据库至关重要。本文阐述了具备通用性的构建下一代临床科研专病数据库的基本理念框架，强调建立多模态数据整合机制、以隐私计算为支撑的多中心协作体系、高智能化的软硬件实体以及一体化的运营管理模式，并介绍这一理念框架在中国乳腺癌创新联盟项目-乳腺癌立体化生命数据库建设中的应用。

**关键词：**专病数据；数据管理；临床科研数据库；数据库建设；乳腺癌
**中图分类号：**R194
**开放科学(资源服务)标识码(OSID):**

## 0　引言

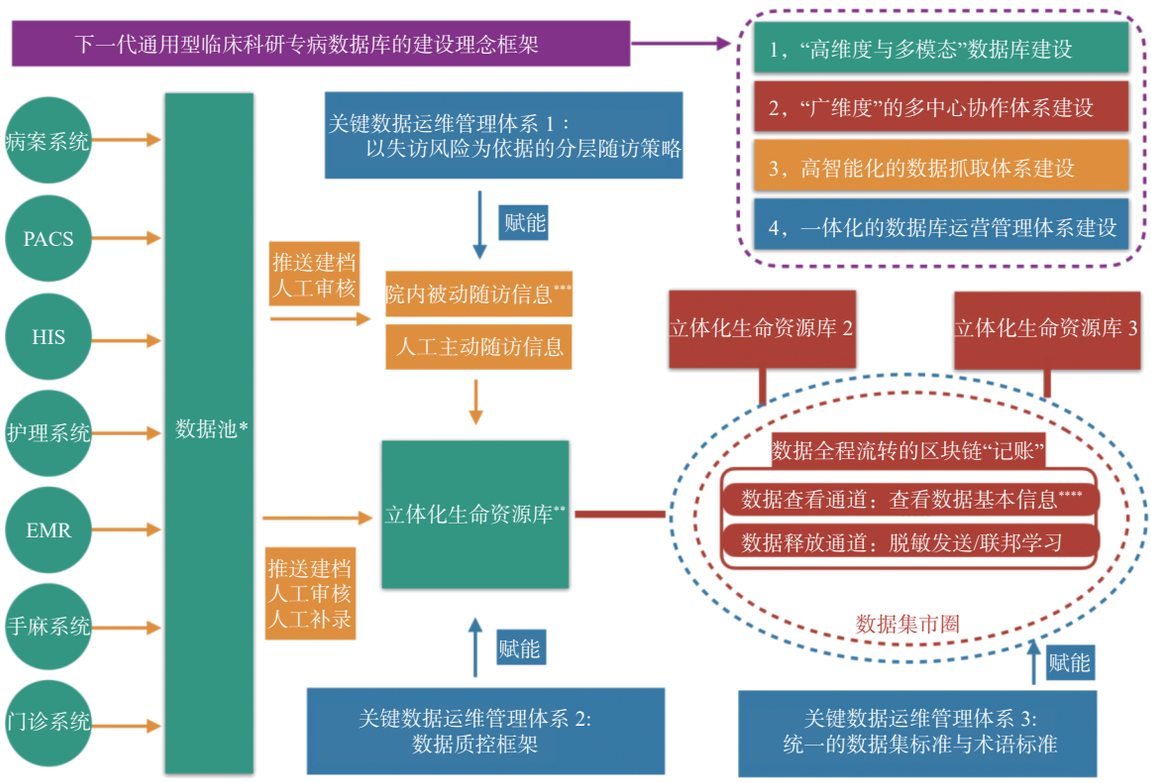
专病数据库的构建基于电子病历数据的筛选、标准化处理及多维度整合，从而形成了结构化且信息丰富的数据集，为专病研究提供了基础支持。这些数据对于临床研究具有重要意义，不仅有助于揭示疾病的发展规律，还能辅助医生总结治疗经验，进而制定更加有效的诊疗方案。此外，通过对专病数据的持续监测和分析，可以识别并改进医疗管理中的问题，从而提高医疗服务的质量。因此，建立专病数据库对于推动临床研究、提升医疗服务质量以及促进医学科学的进步具有显著的重要性。

诊治不同疾病的方法和预后各有差异，因此构建针对特定疾病的数据库时，需要考虑到这些差异性。以ICU病房患者为例，他们的大部分数据都是在病房内生成的，这些数据通常结构化程度高，信息比较完整。然而，对于一些良性疾病，比如只需在门诊进行治疗的病例，可能会面临数据收集不全面以及随访过程中的各种困难。乳腺肿瘤专病数据库的建设也展现出其独特性，首先，乳腺癌患者的生存期相对较长，目前欧美国家和中国乳腺癌的五年生存率分别达到了90%和82%[1-3]。这一长期生存率要求随访周期延长，因此对随访系统的要求更为严格，需要建立一个完善的随访管理体系。其次，乳腺癌的治疗方法多样，包括手术、放疗、化疗、内分泌治疗、靶向治疗、免疫治疗。最后，乳腺癌的诊疗方法更新迅速，新技术和新药物层出不穷，例如基因编辑技术CRISPR-Cas9系统、免疫检查点抑制剂、抗体偶联药物等[4-5]，这就要求专病数据库的数据字段必须具备足够的灵活性和可调整性。

中山大学孙逸仙纪念医院乳腺肿瘤中心的乳腺肿瘤专病数据库建设历程可分为三个阶段：（1）1998—2014年，在此阶段数据库主要依赖于Microsoft Excel和Epidata作为数据载体。这一时期的数据库适用于小样本量、不需要长期纵向随访的医学研究，其优点在于易于获取和操作。然而，从长远角度来看，这些软件在高效、简易的数据管理方面存在局限，数据存储分散，且在数据规模上有所限制；（2）2014—2021年，该阶段引入了美国REDCap eCRF软件[6-7]，REDCap的优势在于支持用户自主建库、便于调整、多用户同时访问和编辑记录、数据修改有追踪记录、随访日历的展示等[8-9]。但REDCap并非专为专病数据库设计、软件源代码未公开，且无法与医院HIS系统进行数据对接（仍需手动录入数据），因此具有一定的局限性。随着医学研究的深入，如单细胞测序、深度学习人工智能等前沿领域[10-12]，REDCap的局限性变得更加明显，尤其是在多模态数据的管理与应用方面。此外，国际局势也使得REDCap用户对系统安全性产生担忧；（3）2021年至今，基于过往的经验和日益增长的临床与科研需求，本团队提出了下一代通用型临床科研专病数据库的建设理念框架（图1），主要内容包括"高维度与多模态"数据库建设、"广维度"的多中心协作体系建设、高智能化的数据抓取体系建设、一体化的数据库运营管理体系建设。这一理念框架可以为临床科研单病种数据库建设与开发提供一定的参考。

## 1　下一代专病数据库需要关注"高维度"与"多模态"

在医学大数据的推动下，虽然各种专病数据库的建设得到了广泛关注和快速发展[13-14]，但目前这些数据库多以扁平化、文本数据为主。扁平化数据，即将复杂的多层数据结构转化为一维结构，虽然简化了数据结构并减少了储存空间，但也带来了数据结构失真和数据完整性受损的问题。以乳腺癌




* 数据池的临床重要性在于保障用户查询下载原始数据不影响临床业务系统;
** 基本的自主数据管理功能包括：1. 完全自主建库管理；2. 自动化建档；3. 智能化CRF管理；4. 智能化随访；5. 自主权限管理
*** 院内被动随访信息的智能化整合是减少人工主动随访工作量的重要步骤
**** 数据基本信息指数据的基本介绍，字段设置方案与数据字典，患者人数以及数完整率，不包含具体的字段信息

**图 1　下一代通用型临床科研专病数据库的建设理念框架体系**

**Figure 1　Conceptual framework of a next-generation general clinical scientific disease database**

临床科研数据库为例，常规数据库记录的TNM分期信息无法全面反映肿瘤的边缘形态、微环境的淋巴细胞浸润模式等对手术和治疗反应性及预后的影响。这就需要应用人工智能、深度学习等高阶技术，对原始图像数据进行深入分析，而传统的扁平化数据库在这方面无法提供有效支持。因此，下一代临床科研专病数据库应当突破传统的数据维度限制，推进多模态数据的整合与资源池的构建。多模态数据通常指以患者为中心，通过不同系统获取的包括文本、图像、语音、视频等多种媒体数据。多模态专病数据库的建设核心在于需要专门的硬件设施来构建数据池，包括：（1）对接各类模态数据（如影像、病理、基因等）业务系统，定期在业务闲时抽取单模态数据，形成各类模态的数据池；（2）以数据记录（患者）为中心，将各类模态数据在时间维度上进行多重整合和统一关联；（3）确保多模态原始数据能够全天候进行便捷的管理、检索和脱敏下载，以支持科研的多维度分析。以早期乳腺癌专病数据库为例，已经整合了包括CT、MR、彩超、钼靶影像数据以及手术病理的高清扫描图、基因测序原始文件等，实现了数据的整理和对齐。这样的整合不仅提高了数据查询的范围和使用效率，也为基于深度学习或人工智能的研究提供了更加丰富、准确的数据支持。

## 2　下一代专病数据库需要广维度的多中心协作体系

专病数据库的广度，即其覆盖范围和能够实现的常态化多中心协作网络的程度，是衡量其效能和影响力的重要指标。中国工程院院士陈赛娟提出：“为提升专病领域的国际竞争力、话语权和国际影响力，建设国家级、标准化、可共享、符合伦理规范、多层次的重大疾病专病数据库，整合临床表型、疾病诊疗信息、临床样本及其生命组学信息等，具有非常重要的战略意义。”美国的SEER数据库是一个成功的案例，它作为一个运营了约半个世纪的多中心多肿瘤专病库，为肿瘤研究提供了宝贵的数据资源[15]。然而，由于国情的差异，我国在医疗行业的多中心协作方面面临着挑战，专病数据库的建设在这一方面仍处于探索阶段。

### 2.1　多中心协作模式

确保每个参与中心都愿意积极分享数据，是多中心协作体系的核心。目前尚无普遍认可的标准方

案，不同医疗机构正积极探索并鼓励彼此间的学习和交流。以乳腺肿瘤多中心协作数据库的建设为例，笔者所在团队参与了两种不同模式的多中心协作数据库建设。第一种模式是由国家卫生健康委医院管理研究所（简称“医管所”）主导的中国乳腺癌标准数据库建设。在这一模式下，各参与医院与医管所签订合作协议，随后在各自医院部署本地服务器用于数据存储。待数据达到质量控制标准后，进行脱敏处理并上传至医管所的国家医院管理研究信息系统，实现数据的整合。第二种模式是由广州市卫生健康委推动的广州市区域乳腺肿瘤防治示范体系（在建）。在这个体系中，笔者所在的中心担任牵头单位，目标是构建一个涵盖筛查、手术、全身治疗、康复等多方面的乳腺肿瘤防治体系。在项目组织的协调下，广州市内的各级医院及社区卫生服务中心的乳腺肿瘤数据得以闭环汇总，形成一个全面、系统、科学的专病数据库。这两种模式各有特点，反映了在多中心协作数据库建设过程中，不同策略和方法的应用。临床上我们鼓励有不同的实践尝试，以更好地理解如何有效地构建和管理多中心协作的专病数据库。

2.2　多中心协作的隐私保护

在多中心协作中，维护数据安全和患者隐私至关重要。专病数据，作为科研的关键部分，受国家数据安全法保护[16-17]。传统上，数据在主中心服务器集中存储和管理，但这可能引起对数据安全的担忧。因此，现采用匿名化处理来保障患者隐私和信息安全。然而，匿名化处理可能导致一些问题，如各中心不愿意本中心的治疗信息（治疗剂量、手术时长等）在匿名后被其他机构分析，存在泄露风险。因此，需要更先进的技术手段加强数据隐私保护。

隐私计算技术，包括多方安全计算、联邦学习和可信安全环境，是跨学科的技术体系，涵盖密码学、人工智能、数据科学等，旨在保障数据隐私下实现共享和计算。目前，虽然医疗领域的隐私计算应用尚不广泛，但国家卫生健康委已推动其在医疗领域的应用，实现保护隐私的数据共享和利用[18]。隐私计算作为保护数据隐私的关键技术，在多中心协作的专病数据管理中具有广阔应用前景。通过在医疗机构内部采用隐私计算进行数据本地分析和处理，再在多中心间共享，既保护了患者隐私和信息安全，又确保了数据使用的合规性和可信度。

2.3　多中心协作的统一数据标准

在医疗领域，多中心协作需要遵循统一的标准，以确保各中心之间的数据和信息能够高效地共享和交换。为此，我中心积极参与并牵头制定了《乳腺癌标准术语集》和《乳腺癌临床科研标准数据集》两套标准[19-20]，推动全国乳腺肿瘤疾病标准体系的建设。通过统一和规范术语集和数据集，为全国范围内的乳腺肿瘤疾病诊疗和研究提供了一套通用的语言和标准化的数据基础，有助于提高诊疗质量和科研水平。同时，也有利于实现跨地区、跨机构的数据共享和信息交流，进一步推动乳腺肿瘤疾病领域的合作和发展。

2.4　多中心协作的贡献评价方式

在多中心协作体系中，客观评估各中心的数据“贡献度”对于激励各方积极、持续参与至关重要。例如，A中心虽提供大量病例数据，但数据质量较低，使用率不高，其“贡献度”相对较低。相反，B中心虽数据量少，但质量高、临床价值大，被广泛采用，其“贡献度”较高。对数据“贡献度”的评价有助于赋予各中心不同的“优先权利”，促进提高数据质量和专业价值。数据“贡献度”的评价可基于数据量、质量、查看次数、提取次数及可信度等。区块链技术在此发挥关键作用，提供一个公开、透明、可追溯的平台，记录数据贡献和利用情况。它确保数据来源和使用途径的透明度，防止不当使用和侵犯利益。同时，区块链还能提供公平的激励机制，鼓励更多中心加入协作。

## 3　下一代专病数据库需要高度智能化的数据提取体系建设

笔者认为，下一代专病数据库应在两方面应用智能化技术：数据自动化抓取与自动化质控。通过自然语言处理（NLP）技术，将非结构化电子病历数据自动转换为结构化专病数据，这被视为理想模式，可显著减少人力。自动化数据抓取分为两类：（1）结构化数据，如年龄、身份证号等，通常可直接、准确、高效地自动抓取，极少需人工审核；（2）非结构化数据，如乳腺肿瘤的TNM分期、病理免疫组织化学指标等，其完整性和准确性依赖于病历书写质量，更多是管理而非技术问题。若病历书写不全或不准确，自动抓取难以实现。现有NLP技术已经可以对非结核化的报告文本如病理报告、影像报告等进行自动后结构化处理以减少工作量，但因不同报告的风格有差异，具体应用时仍需较大量的人力和时间投入进行模型优化。即使如此，该技术在临床中也只能实现约90%的数据录入

准确率，仍需人工审核和质控。

## 4 下一代专病数据库需要一体化的数据运营管理体系建设

下一代临床科研专病数据库的建设需依托一体化的运营管理体系。重视运营管理体系建设意味着不仅注重数据库软硬件功能，更关注整体运作管理。关键措施包括：（1）专人管理，指派专业人员负责数据库管理和维护，保障安全稳定运营，避免医生或研究生承担此类工作；（2）统一培训和关键绩效指标（KPI）考核，定期对数据参与人员进行培训，并通过KPI考核，确保数据准确性和完整性；（3）质控与质量管理，建立严格的质量管理体系，包括数据质量标准、监控和评估，保证数据准确可信；（4）分层随访体系，制定科学随访策略，对高失访风险患者进行密切随访，对普通患者执行常规随访。笔者团队曾提出乳腺癌患者随访失访风险预判模型[8]，已应用于实际运营，提高随访精准性；（5）建立查找失访流程，最大限度减少失访，确保数据完整可用。这些措施共同构建高效、可靠、准确的数据库管理体系，支持单中心乃至多中心协作的数据管理。

## 5 综合应用“理念框架”四项内容提高数据库的提取与利用效率

在临床专病数据库的建设过程中，如何提高数据库的提取与利用效率是一个备受关注的重要问题。笔者认为，建设高维度多模态以及广维度多中心协作的数据库将提高用户使用数据的积极性。提取与利用效率的问题，关键在于提高用户贡献数据的积极性、用户使用数据的积极性和数据库的数据查询与下载的效率。应用本文的理念框架将有助于实现上述目的。处理好整合不同级别和不同机构的数据问题，将有助于提升数据的丰富度、减少重复建设，并促进数据共享。在整合过程中，可以充分利用本框架中提到的几个重要工具，包括：（1）探索合适的合作模式：通过与不同机构建立合作关系（包括参与机制与退出机制），实现数据的共享和整合，从而更好地服务于专病研究；（2）采用统一的数据标准：制定和执行统一的数据标准，确保数据的准确性和一致性，为后续科研数据分析提供可靠的基础；（3）引用隐私计算技术保护数据隐私信息：可以利用隐私计算技术在保护数据隐私的同时进行数据分析，确保数据的安全性和隐私性；（4）运用区块链技术进行数据确权：通过区块链技术，可以明确数据的所有权和使用权，保障数据提供方的合法权益。通过这些工具的应用，可以有效地整合不同级别和（或）不同机构的专病数据，为专病研究提供更为全面和准确的数据支持。

提高数据库的提取与利用效率的技术面支撑离不开高智能化的数据抓取体系建设，以数据记录（患者）为中心，对接各类模态数据（如影像、病理、基因等）业务系统，定期在业务闲时抽取单模态数据，形成各类模态的数据池；以及确保多模态原始数据能够全天候进行便捷的管理、检索和脱敏下载均是有效提升数据查询和下载效率的方式。对提取到的数据进行严格质控，提高数据质量的可靠性，也是吸引用户使用该数据库的重要保证措施之一。在具体实践中，笔者认为对于不同的科研项目的需求，应当需要有不同的质控标准，非一概而论。数据（文字、图像、声音等）的质量应当是合适项目需要为宜，过高的质控标准会提高经济与人力成本，降低数据库的效率，而过低的质控标准则会使数据不可用。

## 6 中国乳腺癌创新联盟项目—乳腺癌立体化生命数据库建设介绍

我国的临床科研专病数据库建设仍处于探索阶段。为实现这个目标，我们也在构建一个多中心、多模态、高效运营的专病数据库，由宋尔卫院士倡导成立的中国乳腺癌产学研创新联盟是一种探索模式，其项目旨在推动资源整合、协同创新，建立以企业为主体、市场为导向、产学研深度融合的创新体系，引领乳腺癌产学研自主创新水平整体跃升，最终提高乳腺癌疾病诊疗水平。为更好地支撑联盟发展，由广州市科技计划项目支持的乳腺癌立体化生命数据库建设是笔者团队参与实践“下一代通用型临床科研专病数据库的建设理念框架”的重要工程项目。截至2023年12月，该项目完成了17 000+乳腺癌患者的结构性数据库建设工作，已接入院内病案首页系统、PACS系统、病理信息系统、护理系统、HIS系统和EMR系统。该项目主要有几个重要的建设内容：

（1）数据池建设：传统的医院在医学信息系统的建设过程中往往会忽视数据池的建设，特别是对于PACS影像系统或病理信息系统的原始数据以往一般存储于生产端/设备端，这不能满足当前人工智能/多模态组学研究范式对数据的需求。为保证用户在查询、提取、下载原始数据的过程中不影响临床工作，有必要建设独立的数据池。乳腺癌立

体化生命数据库将乳腺癌患者相关的7万2千余张高清病理图片，3万2千余张MR原始图像，1万5千余张钼靶，16万余张超声原始DICOM图像，9万3千余张CT原始DICOM图像，1 488份基因原始数据全部单独整理成数据池，为后续的查询和数据提取提供方便。

（2）乳腺癌立体化生命数据库建档随访工作主要依靠专职人员进行维护管理，为进一步提高人员管理的效率，我们采取了以下几种方式：①自动推送建档：系统自动在数据池中抓取动态更新的患者，符合特定条件的患者将被数据库自动抓取后，经人工审核、修正后入档；②数据自动在动态更新的数据池中抓取入档患者，并自动识别是否出现复发转移事件，生成随访信息后，经人工审核修正，记入随访信息；③分层随访体系：针对不同失访风险的患者，采用不同的随访策略。

（3）乳腺癌立体化生命数据库具备完全的用户自主建库管理的能力，用户可自主建设或调整相应的数据专病库，并备有智能化的CRF以及随访管理功能，结合本团队自研的数据质控框架、乳腺癌数据标准集和术语标准集，以最大限度提高该数据库的科学性。

（4）乳腺癌立体化生命数据库的建设采用三种创新技术，以初步探索多中心协作数据集市圈的可能性。如图1示，若干中心同步安装了同一个立体化生命数据库后，通过两个管道互联和一个监控技术，形成了数据集市圈以及新型的数据流转模式：①查看管道：三家中心可查看对方已有数据集的基本信息，包括数据库的基本介绍、数据字段设置方案与数据字典、患者例数以及数据完整率，不包含具体的字段信息；②数据释放管道：数据提供方在接受到数据申请方的数据申请后，可按院内的数据管理流程完成审批，采用两种方式释放数据，一是脱敏后经网络点对点传送释放，二是将数据进入联邦学习系统网络中进行建模，以期做到数据不出院的目的；③监控技术（区块链管道）：多个中心形成的数据集市圈以及上述的数据查看、释放过程，均通过区块链的形式予以记录，确保每一个数据的查看、申请、审批、释放、使用过程全程受监管，保障数据安全的同时，也可以更好地突显数据的价值。

## 7　总结

建设下一代临床科研专病数据库对医院的专科建设至关重要，它不仅在医疗服务和医学研究方面提供显著优势，还有助于吸引更多患者和科研合作，增强医院竞争力。笔者提出的下一代通用型临床科研专病数据库建设理念框架，强调建立一个多模态、多中心协作、高智能化的软/硬件实体，配备一体化的运营管理模式，可有效提升我国临床科研水平，助力医学研究进步，保障公众健康。

**致谢：**

本文获得中山大学孙逸仙纪念医院智慧医疗中心、人工智能研究室、生物医学大数据平台和随访中心支持。

**利益冲突声明：**

所有作者均声明不存在利益冲突。

**参考文献：**

[1] SEER. Cancer Stat Facts: Female Breast Cancer[DB/OL]. 2023. https://seer.cancer.gov/statfacts/html/breast.html.

[2] Cancer Research UK. Breast cancer statistics(2016-2020)[DB/OL]. 2023. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Three.

[3] Xia C, Dong X, Li H, *et al*. Cancer statistics in China and United States, 2022: profiles, trends, and determinants[J]. Chin Med J (Engl), 2022, 135(5): 584-590.

[4] Liu Z, Shi M, Ren Y, *et al*. Recent advances and applications of CRISPR-Cas9 in cancer immunotherapy[J]. Mol Cancer, 2023, 22(1): 35.

[5] Shiravand Y, Khodadadi F, Kashani SMA, *et al*. Immune Checkpoint Inhibitors in Cancer Therapy[J]. Curr Oncol, 2022, 29(5): 3044-3060.

[6] Harris PA, Taylor R, Thielke R, *et al*. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support[J]. J Biomed Inform, 2009, 42(2): 377-381.

[7] Harris PA, Taylor R, Minor BL, *et al*. The REDCap consortium: Building an international community of software platform partners[J]. J Biomed Inform, 2019, 95: 103208.

[8] Ouyang Q, Li S, Gao M, *et al*. Risk factors associated with loss to follow-up of breast cancer patients: A retrospective analysis[J]. Breast, 2021, 57: 36-42.

[9] Yang Y, Liu J, Peng M, *et al*. Introduction of a multicenter online database for non-metastatic breast cancer in China[J]. Sci China Life Sci, 2020, 63(9): 1417-1420.

[10] Chen K, Li H, Pan Z, *et al*. Insights into artificial Intelligence in clinical oncology: opportunities and challenges[J]. Sci China Life Sci, 2022, 65(3): 643-647.

[11] Mustachio LM, Roszik J. Single-Cell Sequencing: Current Applications in Precision Onco-Genomics and Cancer Therapeutics[J]. Cancers (Basel), 2022, 14(3): 657.

[12] 江杨, 于惠芝, 高亚, 等. AI细胞形态学联合DNA定量分析鉴别良恶性胸腹水的探讨[J]. 肿瘤防治研究, 2023, 50(4): 390-396. [Jiang Y, Yu HZ, Gao Y, *et al*. AI Cytomorphology Combined

with DNA-image Cytometry for Identifying Benign and Malignant Pleural Effusion and Ascites[J]. Zhong Liu Fang Zhi Yan Jiu, 2023, 50(4): 390-396.]

[13] 陈可, 王俊. 中晚期非小细胞肺癌专病数据库的构建及应用探讨[J]. 中国数字医学, 2022, 17(6): 109-113. [Chen K, Wang J. Construction and application of the special disease database for middle-advanced non-small cell lung cancer[J]. Zhongguo Shu Zi Yi Xue, 2022, 17(6): 109-113.]

[14] 陈嘉旖, 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. [Chen JY, Tao R. Construction of the Special Disease Database of Lung Cancer Based on Clinical Big Data[J]. Zhongguo Shu Zi Yi Xue, 2021, 16(11): 39-42.]

[15] Siegel RL, Miller KD, Fuchs HE, *et al*. Cancer statistics, 2022[J]. CA Cancer J Clin, 2022, 72(1): 7-33.

[16] 张宇.《数据安全法》中规定的法律义务[J]. 中国信息化, 2023(10): 36-38. [Zhang Y. Legal Obligations under the Data Security Law of China[J]. Zhongguo Xin Xi Hua, 2023(10): 36-38.]

[17] 何志鹏, 朱志远. 国家安全法体系的边界[J]. 山东大学学报(哲学社会科学版), 2023, (6): 118-128. [He ZP, Zhu ZY. The Boundary of the National Security Legal System[J]. Shandong Da Xue Xue Bao (Zhe Xue She Hui Ke Xue Ban), 2023, (6): 118-128.]

[18] 国家卫生健康委员会. “十四五”全民健康信息化规划[DB/OL]. 2022. http://www.nhc.gov.cn/cms-search/xxgk/getManuscriptXxgk.htm?id=49eb570ca79a42f688f9efac42e3c0f1. [National Health Commission of the People’S Republic of China. The 14th Five-Year Plan for National Health Informatization[DB/OL]. 2022. http://www.nhc.gov.cn/cms-search/xxgk/getManuscriptXxgk.htm?id=49eb570ca79a42f688f9efac42e3c0f1.]

[19] 中国临床肿瘤学会. 乳腺癌临床科研标准数据集: T/CSCO 2—2022[S/OL]. 2022. [Chinese Society of Clinical Oncology. Standard Datasets for Breast Cancer Clinical Scientific Research: T/CSCO 2—2022[S/OL]. 2022.]

[20] 中国临床肿瘤学会. 乳腺癌名词术语: T/CSCO 1—2022[S/OL]. 2022. https://www.ttbz.org.cn/Pdfs/Index/?ftype=st&pms=66386. [Chinese Society of Clinical Oncology. Terminology of Breast Cancer: T/CSCO 1—2022[S/OL]. 2022. https://www.ttbz.org.cn/Pdfs/Index/?ftype=st&pms=66386.]

[**编辑**：尤婷婷；**校对**：邱颖慧]

**作者贡献：**

**彭　敏**：文章撰写

**陈倩茹**：作图及文章修改

**陈　凯**：文章审阅及修改