ICS 35.240.80
C 07

团 体 标 准

T/BSIA 029.1-2026

多模态医疗科研数据平台 第 1 部分：总体要求

Multimodal medical research data platform Part 1: general requirements

2026-01-22 发布　　2026-01-23 实施

北京软件和信息服务业协会　发布

目　次

前　言

《多模态医疗科研数据平台》已发布 3 个部分：

——第 1 部分：总体要求；

——第 2 部分：平台架构；

——第 3 部分：科研数据集；

本文件是《多模态医疗科研数据平台》的第 1 部分。

本文件依据GB/T1.1-2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》给出的规则起草。

本文件由中国人民解放军总医院提出，北京软件和信息服务业协会归口。

本文件起草单位：中国人民解放军总医院、医疗大数据应用技术国家工程研究中心、中国人民解放军火箭军总医院、国家人口健康科学数据中心、北京大学、浙江大学、中南大学、北京嘉和美康信息技术有限公司、北京嘉和海森健康科技有限公司、万达信息股份有限公司、北京软件和信息服务业协会。

本文件主要起草人：何昆仑、石金龙、乌日力格、陈媛媛、王万玲、吴欢、车贺宾、卢敬泰、周伟、黄正行、黄雨、龙军、陈先来、肖筱华、张敬晨、闫雯、窦飞、陈鹏、龙飞、张磊、仓剑。

引　言

本标准旨在更好推广相关国家重点项目成果，建设应用医疗科研数据共享平台，整体内容计划分为 6 个部分：

——第 1 部分：总体要求；
——第 2 部分：平台架构；
——第 3 部分：科研数据集；
——第 4 部分：样本数据收集、存储及传递；
——第 5 部分：知识工程；
——第 6 部分：共享管理。

这 6 个部分共同组成完整的多模态医疗科研数据平台标准体系：

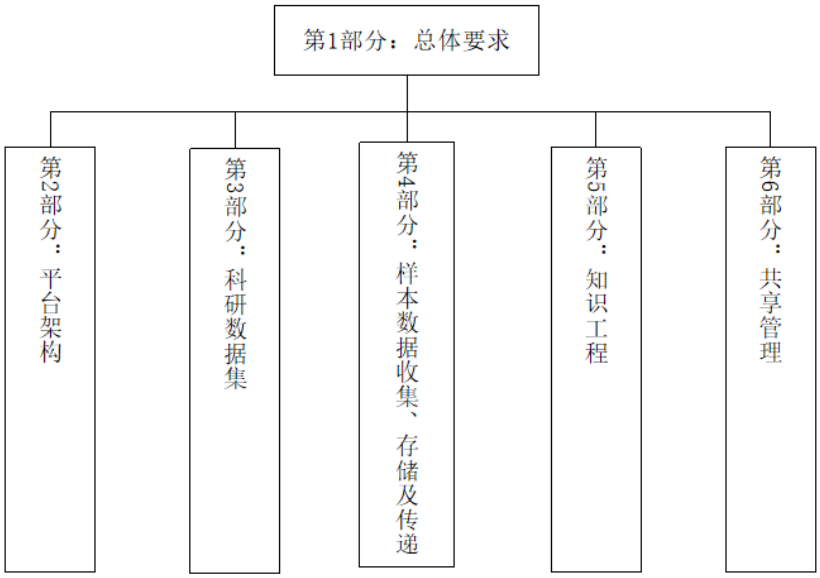


图 1　多模态医疗科研数据平台总体构成

第 1 部分：总体要求。规定平台建设的总体框架和通用要求，定义标准体系结构和各组成部分间的关系。

第 2 部分：平台架构。规定平台技术架构和功能模块，定义平台数据的接入、存储、处理和服务的技术规范。

第 3 部分：科研数据集。规定平台使用的疾病科研应用相关的核心数据元、数据集结构和数据格式标准。

第 4 部分：样本数据收集、存储及传递。规范疾病、实验样本数据收集、存储、传递等过程的处理方法，约定样本数据统一编号、命名规则、数据包及公共元数据、数据文件存储目录及传递接口协议等。

第 5 部分：知识工程。将分散的多模态数据转化为结构化的知识体系，支撑智能诊疗、临床决策辅助等高级应用。

第 6 部分：共享管理。规定数据共享使用中的角色管理、安全保护和隐私保护要求，定义访问控制、审计追踪和合规性管理机制。

各部分之间的关系：第 1 部分规定多模态医疗科研数据平台的总体框架和通用要求；第 2 部分为后续各部分提供技术支撑；第 3 部分为第 4 部分提供源数据的规范样例；第 4 部分

完成数据到信息的转化，为第 5 部分知识工程提供高质量数据输入；第 5 部分知识工程实现信息到知识的升华，第 6 部分实现知识输出，为各医疗科研机构提供支撑。各部分既相对独立又有机统一，共同构成完整的标准体系，支撑多模态医疗科研数据的全生命周期管理和应用。

多模态医疗科研数据平台

第 1 部分：总体要求

1 范围

本文件规定了多模态医疗科研数据平台的总体构成和通用规则。

本文件适用于医疗科研用户（医学科研教学单位、医院等）的多模态科研数据应用、医疗行业软件企业和医疗数据处理服务企业或机构的多模态医疗科研数据相关系统开发和应用。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 42135—2022 智能制造 多模态数据融合技术要求 3 术语和定义

GB/T 44109—2024 信息技术 大数据 数据治理实施指南 3 术语和定义

3 术语和定义

《常用临床医学名词（2023年版）》和WS/T445-2014中包含的术语适用于本文件，本文件中仅列出未包含的术语。

3.1

多模态数据

多种形态的数据。

注：包含结构化数据（例如业务系统数据等）、半结构化数据（例如 XML 文件、JSON 文件等）和非结构化数据（例如文本、语音和图像视频等）。

[来源：GB/T 42135—2022]

3.2

数据治理

对数据资源管理行使权力和控制的活动集合(计划、监督和执行)。

[来源：GB/T 44109—2024]

3.3

数据融合

将同一对象的多维度数据进行汇总，从而对该对象形成综合认知。

[来源：GB/T 42135—2022]

3.4

文档型数据库

一种非结构化数据管理系统，其核心特征是以文档为基本单位进行数据的存储、管理和查询。文档通常采用 JSON、BSON 或 XML 等格式，允许每个文档拥有独立、可变的结构（模式）。

3.5

分布式全文检索引擎

一种能够在分布式集群环境中，对文本数据建立倒排索引并提供近实时的、高性能的全文检服务的技术平台。其核心功能包括海量数据的分布式存储、索引、检索和分析。

3.6

知识图谱

以结构化形式描述的知识元素及其联系的集合。

[来源：GB/T 42131-2022]

4 缩略语

《常用临床医学名词（2023年版）》，包含的缩略语适用于本文件，本文件中仅列出未包含的缩略语。

下列缩略语适用于本文件：

BSON 一种二进制编码格式，用于表示类似 JSON 的文档（Binary JSON）

EMR 电子病历系统（Electronic Medical Record）

DICOM 医学影像存储与传输标准(Digital Imaging and Communications in Medicine)

HIS 医院信息系统（Hospital Information System）

JSON JavaScript 对象表示法（JavaScript Object Notation）

LIS 检验信息系统（Laboratory Information System）

PACS 影像归档和通信系统（Picture Archiving and Communication System）

RBAC 基于角色的访问控制（Role-Based Access Control）

5 总体构成

5.1 平台架构

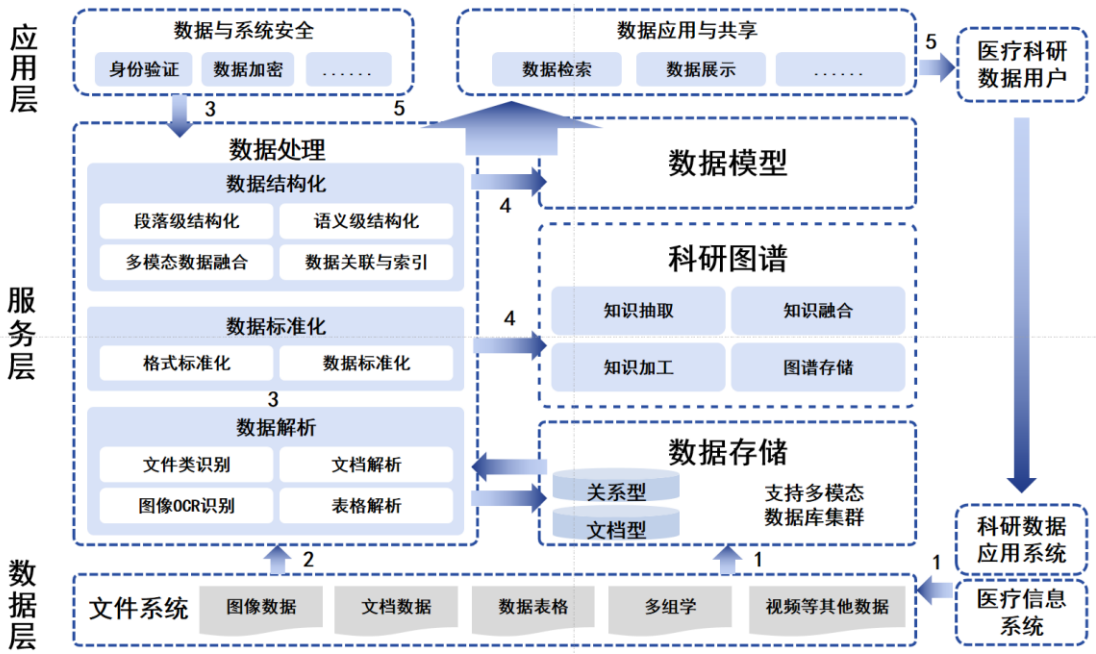
本平台采用分层模块化的逻辑架构设计，以实现高内聚、低耦合的系统组织方式，确保系统的可扩展性、可维护性与安全性。整体逻辑架构分为数据层、服务层与应用层三层结构。

数据层作为系统基础，负责多模态科研数据的存储与管理，包括结构化、半结构化和非结构化数据。该层采用混合存储策略，集成文档型数据库与分布式全文检索引擎，分别实现非结构化数据的灵活存储与高效检索功能。

服务层是系统的核心处理单元，提供数据处理、知识图谱构建与管理、自然语言处理、多模态语义对齐等核心服务。该层通过微服务架构实现功能模块化，各服务可独立部署、扩展和升级，包括数据解析引擎、自然语言分词与结构化引擎、知识获取与融合引擎、数据分析引擎等。

应用层面向最终用户，提供直观易用的操作界面，包括系统登录、数据检索、项目管理、数据分析、个人中心等功能模块。该层通过统一的API网关与服务层交互，确保用户操作的响应速度与系统安全性。

逻辑结构如图1所示，其中数字标号表示数据流程节点，详见5.2。



各层之间通过标准化接口进行通信，采用RESTful API与消息队列等多种方式，保证数据传输的可靠性、实时性与安全性。系统还设有统一的安全认证与权限管理机制，贯穿所有层次，确保数据与操作的安全可控。

系统采用混合型数据存储架构，以适应多模态医疗科研多模态数据的多样化存储与高效检索需求。

文档型数据库存储用于管理非结构化与半结构化数据。文档型数据库基于JSON的非关系型数据模型，使用BSON（Binary JSON）格式存储文档，适用于存储文献全文、图像、表格、视频、音频等异构科研数据。其文档模型支持动态字段扩展，便于直接映射多模态数据解析后的结构化结果。通过数据规范化与合理的文档结构设计，实现数据的高效存储与查询。系统支持自动索引优化，根据查询模式自动创建索引，提升查询性能。

分布式全文检索引擎专门用于实现高效、实时的数据检索功能。其倒排索引技术支持全文检索、结构化检索与多条件组合检索，满足科研用户对疾病名称、研究项目、关键指标等内容的快速检索需求。分布式架构支持水平扩展，通过分片与副本机制保障高可用性与数据可靠性。分布式全文检索引擎支持多租户管理，可实现不同研究项目或用户组的数据隔离与安全访问。

此外，系统通过数据输入与连接管理模块，实现多模态数据的安全接入与存储。数据输入阶段，系统接收经过解析与标准化处理的结构化数据，将其转换为JSON等适用格式。数据库连接采用加密传输与身份验证机制，确保数据安全。在存储过程中，系统执行数据校验与异常处理，保障数据完整性与一致性。

5.2 数据流程

多模态医疗科研数据应用的总体流程应遵循“数据采集与存储→数据管理与质量控制→数据整合与脱敏→数据分析与知识发现→安全共享与价值反馈”的全生命周期管理理念，该流程体现了数据从原始状态到知识化应用的完整转化过程，各环节相互衔接，构成数据治理活动的核心环节。

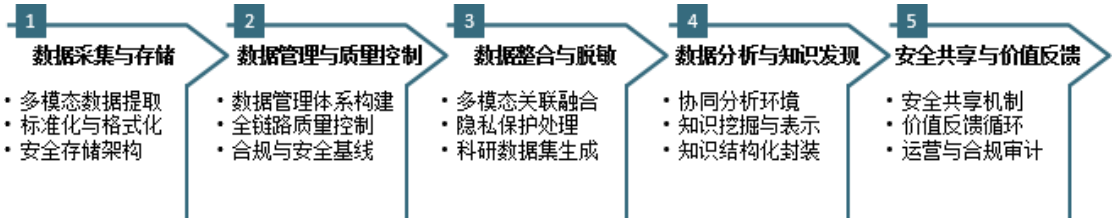


图 2　多模态医疗科研数据平台的总体流程

5.2.1　数据采集与存储，实现多源异构数据的标准化接入与安全存储，包括：

a）　多模态数据提取：从 HIS、EMR、LIS 等系统提取临床诊疗数据（结构化数据）；从 PACS 系统提取 DICOM 格式的影像数据；从组学平台提取基因组、蛋白质组等生物数据；从物联网设备提取连续监测的时序数据。

b）　标准化与格式化：采用相关标准转换临床数据，确保接口交互规范性；使用标准的临床术语系统、实验室检验和临床观察项目编码标准等标准术语以统一语义；组学数据遵循基因组数据共享和分析的相关开放标准。

c）　安全存储架构：结构化数据存入关系型数据库；非结构化数据（影像、文本）存入对象存储或文档数据库；建立分布式检索集群支持高效查询。

5.2.2　数据管理与质量控制，建立体系化治理框架，确保数据可信可用，包括：

a）　数据管理体系构建：元数据管理，定义技术元数据（字段类型、来源）、业务元数据（指标含义）、管理元数据（血缘关系），形成数据资产目录；数据标准管理，制定统一的数据元标准（如字段长度、值域约束），覆盖临床、影像、组学等全维度；主数据管理，确保核心实体（患者、药品、科室）跨系统一致性。

b）　全链路质量控制：定义质量规则（完整性、准确性、时效性）；执行自动化数据剖析（缺失值检测、异常值清洗）；生成质量评估报告，驱动源头系统改进。

c）　合规与安全基线：分类分级，按敏感程度标记数据（如个人信息、基因数据为受限级）；访问控制，基于 RBAC 模型（角色权限控制）实现字段级权限管理。

5.2.3　数据整合与脱敏，在隐私保护前提下，形成科研数据集，包括：

a）　多模态关联融合：通过可信主索引关联同一患者的临床、影像、组学数据；构建时空对齐模型。

b）　隐私保护处理：匿名化，删除直接标识符（姓名、身份证号），泛化间接标识符（年龄分组、地理区域模糊）；假名化，采用令牌化技术替换标识符，支持受控重标识；严格遵循《个人信息保护法》和 GB/T 37973-2019。

c）　科研数据集生成：按研究课题需求，从数据湖中提取、筛选样本队列；输出包含特征字段、标签字段的标准分析数据集。

5.2.4　数据分析与知识发现，从数据中提炼洞察，形成可复用的科研知识资产，包括：

a）　协同分析环境：提供封装的分析工具；集成机器学习库和组学分析流程。

b）　知识挖掘与表示：应用人工智能技术解析临床文本，提取实体关系（疾病-症状-药物）；通过影像 AI 算法量化病灶特征（肿瘤体积、纹理参数）；多组学整合分析识别生物标志物组合。

c）　知识结构化封装：将分析结果转化为标准化知识对象。

5.2.5　安全共享与价值反馈，实现数据价值安全流转，形成闭环优化，包括：

a） 安全共享机制：联邦学习，跨机构协作建模（数据不出域）；可信执行环境（TEE），提供加密计算沙箱；API 网关，授权访问脱敏数据集或知识服务。
b） 价值反馈循环：研究成果（如疾病预测模型）反馈至临床系统，辅助诊断决策；共享过程中的数据使用日志和质量问题，反向驱动治理规则优化（如补充缺失字段、修订术语映射）。
c） 运营与合规审计：制定数据使用协议和利益分配机制；全操作链审计追踪，满足合规要求。

6 通用规则

6.1 数据应用范围

多模态医疗科研数据应用涵盖从数据采集到知识应用的全过程，具体包括以下范围：

6.1.1 数据来源范围

临床诊疗数据：包括电子病历（EMR）、检验检查结果、医学影像、病理报告等；
生物样本数据：源自血液、组织等生物样本的基因、转录、DNA 甲基化、蛋白质、代谢、微生物等多组学数据；
随访观察数据：患者出院后的随访记录、生活质量评估等长期观察数据；
物联网数据：基于穿戴式设备等物联设备获得连续性生理监测数据；
科研项目数据：各类专项研究产生的专题数据集，如心脑血管疾病研究。

6.1.2 数据类型范围

结构化数据：数据库表格、标准化编码数据等；
半结构化数据：XML、JSON 等格式的科研数据；
非结构化数据：医学影像文件、文本报告、音频记录、组学序列数据文件等；
时序数据：连续监测的生命体征、长期随访数据等。

6.1.3 数据层级范围

原始数据：直接从源系统采集的未加工数据；
处理后数据：经过清洗、转换、标准化处理的数据；
知识化数据：通过分析挖掘形成的知识图谱、模型参数等。

6.2 平台操作流程

6.2.1 前台科研操作流程，科研人员在前台的主要数据处理活动包括：

a） 数据申请流程
需求提出:科研人员通过平台提交数据使用申请,明确研究目的、数据范围和使用期限；
监管审查：申请提交监管方进行科学性、伦理性审查；
权限授予：审查通过后，系统自动分配相应数据访问权限；
使用监督：平台记录数据使用全过程，确保合规使用。
b） 数据导入流程
数据准备：科研人员按照平台要求的数据格式准备研究数据；
质量自查：对准备导入的数据进行初步质量检查；

上传提交：通过安全通道上传数据至平台临时存储区；
系统验证：平台对上传数据进行格式验证和完整性检查。
c） 数据使用流程
数据探索：通过平台提供的工具进行数据浏览和探索性分析；
分析建模：利用平台分析工具进行统计分析和模型构建；
结果导出：经审批后导出分析结果（不含原始敏感数据）；
成果归档：将研究过程和成果归档至平台知识库。

6.2.2 后台管理流程，数据管理人员在后台执行的核心数据处理活动：

a） 多模态数据处理
数据接收：从各源系统接收多模态数据，建立数据溯源机制；
质量评估：对接收数据进行全面的质量评估，包括完整性、准确性、一致性等指标；
标准映射：将源数据映射到标准医学术语体系；
数据增强：通过数据清洗、转换、增值等操作提升数据质量。
b） 多模态数据解析
结构化数据解析：对数据库表格等结构化数据进行解析和标准化；
文本数据解析：采用自然语言处理技术解析临床文本报告；
影像数据解析：对医学影像数据进行特征提取和标准化描述；
时序数据解析：对连续监测数据进行分段、滤波和特征提取。
c） 自然语言分词及结构化
文本预处理：对临床文本进行分词、去噪、标准化等预处理；
实体识别：采用深度学习技术识别医疗实体（疾病、药物、手术等）；
关系抽取：提取实体间的语义关系（治疗、诊断、不良反应等）；
结构化存储：将提取的结构化信息存入知识库。
d） 多模态语义对齐
特征对齐：将不同模态数据的特征映射到统一的语义空间；
关联建立：建立跨模态数据间的语义关联关系；
一致性验证：确保多模态数据在语义层面的一致性；
知识融合：将多源异构数据融合成 统一的知识表示。

6.2.3 端到端流程整合，建立前后台协同的数据处理流水线，确保数据从采集到应用的完整闭环：前台需求驱动后台数据处理优先级，后台治理成果直接支撑前台科研应用，建立双向反馈机制，持续优化数据处理流程，实现数据处理全过程的可视化监控和质量管理。

6.3 数据共享要求

围绕合规、受控、可溯的核心目标，其要点如下：

a） 分级分类共享：依据数据敏感度（如个人信息、基因数据）进行分级（公开、协议、非公开），制定差异化的共享策略与访问权限。
b） 权限受控申请：建立统一的线上数据申请与审批流程，基于 RBAC 模型进行精细授权，确保数据“按需使用、最小授权”。
c） 安全技术保障：共享过程需采用联邦学习、可信执行环境或 API 网关等安全技术，实现数据“可用不可见”，严防原始数据泄露。
d） 全程审计溯源：记录数据申请、审批、访问、使用的全链条操作日志，支持事后审计与责任追溯，满足合规要求。

参考文献

[1] GB/T 37973-2019 信息安全技术 大数据安全管理指南
[2] WS 445.1—2014 电子病历基本数据集 第1部分：病历概要