ICS 35.240
CCS A01

团 体 标 准

T/BSIA 00X-2025

医疗健康领域衍生数据认定标准

Criteria for Identification of Derived Data in the Healthcare Field

（征求意见稿）

2024-XX-XX 发布　　2024-XX-XX 实施

北京软件和信息服务业协会　发布

目　次

前　　言

本文件按照GB/T 1.1—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》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由北京软件和信息服务业协会提出并归口。

本文件起草单位：北京软件和信息服务业协会

本文件主要起草人：

本文件为首次发布。

引　　言

为进一步规范医疗健康领域衍生数据的认定流程，明确数据权属边界与安全管理要求，推动衍生数据在医疗科研、临床服务、产业创新中的合规利用，特制订本文件。

本文件依据《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《医疗数据安全指南》（GB/T 42439-2023）《健康医疗大数据标准、安全和服务管理办法（试行）》（国卫规划发〔2018〕23 号）及《关于加强医疗机构数据安全管理的通知》等政策法规与标准规范，结合医疗健康数据全生命周期管理实践、衍生数据技术处理趋势，以及企业协会在数据合规服务中的成熟经验，对医疗健康领域衍生数据的认定原则、界定范围、认定流程等核心认定维度提出明确要求，并构建衍生数据的标准化认定流程与评估规范。

本文件由行业协会牵头，联合医疗机构、数据流通企业共同制定，既响应医疗健康产业对数据价值挖掘的需求，也契合数据安全与隐私保护的监管要求，保障衍生数据合规流通与安全应用，为医疗健康领域数据要素市场化配置与产业创新发展提供支撑。

医疗健康领域衍生数据认定标准

1 范围

本标准规定了医疗健康领域衍生数据的术语和定义、界定范围、来源、伦理要求、监督机制以及数据全生命周期管理等内容。适用于指导医疗健康领域衍生数据的认定工作。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 25069-2022 信息技术 安全技术

GB/T 39725 信息安全技术 健康医疗数据安全指南

GB/T 38667 信息技术 大数据 数据分类指南

GB/T 43697-2024 数据安全技术 数据分类分级规则

3 术语和定义

下列术语和定义适用于本文件。

3.1

医疗健康领域原始数据 Raw data in the healthcare field

直接来源于医疗健康相关活动的初始数据，反映个体健康状态、诊疗过程、医学研究结果或群体健康特征，包括但不限于电子病历、检验检查报告、医学影像数据、病理切片数据、用药记录、生命体征监测数据、基因测序数据、公共卫生事件报告数据等。

3.2

医疗健康领域衍生数据 Derived data in the healthcare field

指数据处理者对其依法享有使用权的医疗健康原始数据，通过专业技术加工(清洗、标准化、集成、分析、挖掘、聚合、重构等)、建模分析、特征提取等手段，实现内容、形式或结构实质性改变后形成的具有更高应用价值的数据产物（如疾病预测模型、健康风险评分、医学知识图谱等）。

医疗健康领域衍生数据不包括：(1) 原始数据的简单格式转换(2) 数据存储介质的物理迁移(3) 未改变数据本质的加密/解密过程。

3.3

数据处理者 Data processor

指依法开展医疗健康原始数据收集、存储、加工、分析、传输等活动，并生成衍生数据的组织或机构，包括医疗机构、高等院校、科研院所、医疗科技企业、公共卫生机构等。

3.4

临床检验数据 Clinical test data

指由医学检验科、医学实验室、放射科等产生的针对患者体液、组织、器官的分析检测结果，包括生化检验、免疫检验、微生物学检验、分子生物学检查、影像检查报告等。

3.5

日常健康数据 Daily health data

包括个人健康管理、随访监测、智能穿戴设备采集的生理参数、互联网医院平台线上问诊及自测数据、康复训练数据等，以及患者主动录入的健康管理信息。

3.6

医疗项目数据 Medical project data

涵盖手术、诊疗操作、专科检查、康复治疗、药物管理、用药方案、治疗过程记录、远程医疗交互数据，以及疾病管理、临床路径、科研项目等产生的数据流，含费用、医保相关信息。

3.7

数据主体 Data subject

指医疗健康原始数据所指向的自然人，即患者、健康体检者、科研参与者等数据提供个体。

3.8

衍生数据生成 Derived data generation

指数据处理者通过清洗、转换、建模、分析、整合等技术手段，将医疗健康原始数据转化为衍生数据的过程

3.9

去标识化 De-identification

指通过对原始数据进行处理，使其不能直接或间接识别特定数据主体的过程，且处理后的数据在不借助额外信息的情况下无法恢复识别。

3.10

匿名化 Anonymization

指通过对原始数据进行处理，使其无法识别特定数据主体且不能复原的过程，匿名化后的数据不属于个人信息。

3.11

数据融合 Data fusion

指通过多源数据整合形成新数据集的过程，需注明来源与融合规则。

3.12

数据权属链 Data ownership chain

衍生数据生成过程中涉及的原始数据使用权、加工权、收益权的完整凭证记录。

4 认定原则

4.1 合法性原则

医疗健康领域衍生数据的产生、处理和使用必须符合《个人信息保护法》《数据安全法》等国家法律法规及行业监管要求，原始数据的获取需具备合法来源（如患者知情同意、机构合规授权等），数据加工过程不得违反数据主权与权属相关规定。

4.2 关联性原则

医疗健康领域的衍生数据与原始数据需存在明确的逻辑关联，其产生过程可追溯。应完整记录原始数据来源、加工方法、处理步骤等关键信息，确保衍生数据能够回溯至对应的原始数据，且数据转化过程可解释、可验证。应通过哈希值、区块链存证等技术确保溯源记录不可篡改，原始数据与衍生数据的关联验证响应时间应≤5 秒。

4.3 准确性原则

医疗健康领域衍生数据应真实、客观地反映原始数据的核心特征，不得因加工处理导致信息失真。数据清洗、集成、分析等环节需采用科学方法，对异常值、缺失值的处理需有明确标准，确保衍生数据的精度满足预期应用场景需求。

4.4 安全性原则

医疗健康领域衍生数据的全生命周期（采集、存储、传输、使用、销毁）需采取安全防护措施，包括数据加密、访问权限管控、脱敏处理等，防止数据泄露、篡改或滥用，尤其需保护涉及个人隐私的敏感信息。

4.5 实用性原则

医疗健康领域衍生数据应具备明确的应用场景和业务价值，如辅助临床决策、支持科研创新、优化健康管理等。其结构与格式需符合行业通用标准，便于跨机构、跨系统的共享与复用，避免无意义的数据加工与冗余产出。

5 医疗健康领域原始数据的界定范围

5.1 原始数据来源

5.1.1 患者病例数据

包括患者基本信息、病史、诊断结果、治疗方案、病程记录等，来源于医疗机构的电子病历系统、纸质病历数字化文件等。

5.1.2 临床检验数据

涵盖实验室检验结果、影像学检查报告、病理检查结果等，来源于医学检验科、影像科、病理科等科室的检测设备和信息系统。

5.1.3 日常健康数据

包含智能穿戴设备采集的心率、步数、睡眠数据，家庭健康监测设备记录的血压、血糖数据，以及患者通过健康 APP 上报的症状、生活方式等信息。

5.1.4 医疗项目数据

涉及手术操作记录、药物使用记录、康复治疗过程数据、医疗费用数据等，来源于医院的住院管理系统、门诊管理系统、药房管理系统等。

5.2 原始数据核心特征

5.2.1 直接性

直接来源于医疗健康活动的第一手数据，未经加工、整合或分析，是数据流转的初始形态，如检验设备直接输出的原始测量值而非经统计后的均值。

5.2.2 原始性

保留数据产生时的原始状态，包括数据的格式、精度、记录方式等，未经过标准化转换、错误修正或逻辑重组，可能存在重复、缺失或格式不统一等情况。原始数据应保留以下元数据：数据生成时间戳、采集设备/系统标识码、操作者身份标识。

5.2.3 基础性

基础性是衍生数据产生的源头和基础，所有医疗健康衍生数据均需以原始数据为加工对象，其质量直接影响衍生数据的可靠性与价值。

6 医疗健康领域衍生数据界定范围

6.1 衍生数据分类

6.1.1 标准化衍生数据

基于原始数据进行格式统一、错误修正和规则适配后形成的标准化成果，是数据流转与复用的基础。这类数据包括将不同系统、不同设备产生的非统一格式数据转换为结构化格式，统一字段命名、数据单位和时间格式；通过自动化工具与人工审核结合，剔除重复记录、修正逻辑错误、填补关键缺失值，确保数据的完整性与一致性；以及将原始数据中的本地编码映射至国际或国家通用编码体系，实现跨机构数据的语义互通。

6.1.2 数据脱敏衍生数据

通过数据清洗、去标识化、匿名化等严格的隐私保护加工方式处理后，形成的不包含可识别个人信息的衍生数据，旨在平衡数据利用价值与隐私安全。这类数据的加工过程包括对原始数据中的直接标识符进行彻底删除或替换，对间接标识符进行模糊化处理，同时结合差分隐私技术在数据集中添加适量噪声，确保无法通过单个或多个字段的组合追溯到具体个体。此类衍生数据虽剥离了个人信息属性，但仍保留了医疗健康数据的统计分析价值，可安全应用于医学科研、公共卫生研究、政策制定等场景，有效降低数据共享与复用中的隐私泄露风险。

6.1.3 多源整合衍生数据

整合跨场景、跨时间、跨系统的原始数据，形成覆盖个体全周期或特定主题的综合性数据集。具体涵盖关联同一患者在不同医疗机构的各类诊疗记录、体检报告，以及家庭健康监测设备记录、互联网问诊记录等，构建贯穿全生命周期的健康数据集；围绕特定疾病，整合患者的诊断信息、用药记录、检验结果、影像资料、手术记录、随访数据等多维度信息，支持专病诊疗与研究；通过区域医疗信息平台，整合不同层级医疗机构的原始数据，形成区域化健康数据集，用于分级诊疗与公共卫生服务。

6.1.4 深度分析衍生数据

通过专业算法或深度挖掘技术，从原始数据中提取隐藏信息或生成预测性结果，具有高附加值应用价值。涵盖基于机器学习、深度学习等算法对原始数据训练后产生的各类输出结果、从非结构化原始数据中提取的可量化特征以及通过关联分析算法挖掘的原始数据中隐藏的关联关系等。这类数据是对原始数据的深度解读与价值挖掘，能够为医疗健康领域的决策提供深层次的信息支持。

使用生成式AI加工的衍生数据需：

a) 标注模型名称及版本；

b) 声明训练数据来源；

c) 提供可解释性报告（含SHAP值/LIME分析）

6.1.5 统计汇总衍生数据

对个体原始数据进行群体层面的汇总与统计，形成反映群体特征或行业状况的指标性数据。包含按地域、时间、人群特征等维度聚合生成的公共卫生统计指标；基于医疗机构原始数据计算的医疗质量指标数据，用于医疗质量监管与改进；以及整合多机构原始数据，生成反映医疗行业状况的统计结果，为政策制定提供依据。

6.2 衍生数据核心特征

6.2.1 价值增值性

通过专业加工实现数据价值的显著提升，不仅保留原始数据的核心信息，还通过标准化、整合、分析等手段挖掘出隐藏价值，如从零散的检验数据中提炼出疾病风险规律，从多源病历中整合出个性化治疗依据，其应用价值远超原始数据的简单叠加。

6.2.2 形态转化性

在数据内容、形式或结构上发生实质改变，并非原始数据的直接复制或轻微调整。如非结构化的病历文本转化为结构化的诊断标签，孤立的单源数据整合为多维度的主题数据集，静态的原始记录生成为动态的预测模型输出。

6.2.3 可追溯性

衍生数据与原始数据应存在明确的逻辑关联和溯源路径，能够通过完整的加工记录回溯至对应的原始数据，且转化过程具有可解释性，确保衍生数据的可靠性可验证，避免无依据的数据创造。衍生数据生成应记录原始数据哈希值、加工算法git commit ID、运行环境Docker镜像SHA256码。

6.2.4 场景适配性

针对特定应用场景设计加工，其结构、格式及内容均与临床诊疗、科研创新、健康管理等场景的需求高度匹配。

6.2.5 安全合规性

在加工与应用过程中嵌入安全设计，尤其是涉及个人健康信息的衍生数据，需通过脱敏、匿名化等处理剥离直接或间接标识，既满足数据利用需求，又符合隐私保护与数据安全相关法规。

7 认定流程

7.1 原始数据溯源确认

对拟加工为衍生数据的原始数据进行来源核查，明确原始数据的产生主体、采集场景、数据类型及获取方式，确认原始数据的合法性与完整性，确保其符合相关法律法规及行业规范要求，为衍生数据的认定奠定合规基础。

7.2 加工过程记录审核

核查衍生数据从原始数据到最终形成的全流程加工记录，包括所采用的加工方法（如清洗、标准化、集成、分析等）、技术工具、操作步骤、参数设置及各环节处理结果等，确保加工过程可追溯、可验证，且符合数据处理的科学规范。

7.3 数据属性判定

依据本标准对衍生数据的界定范围，判定待认定数据所属的类别（如标准化衍生数据、多源整合衍生数据、深度分析衍生数据、统计汇总衍生数据、数据脱敏衍生数据等），明确其数据形态、加工深度及应用属性，为后续的合规性与适用性评估提供依据。

7.4 合规性与安全性评估

评估衍生数据是否符合《个人信息保护法》《数据安全法》等相关法律法规要求，重点检查涉及隐私保护的衍生数据是否采用了必要的去标识化、匿名化等处理措施，去标识化处理后的数据重识别风险应<0.1%（采用NIST SP 800-188评估方法）；同时评估数据的安全性，包括数据存储、传输、使用等环节的安全防护措施是否到位，是否存在数据泄露、篡改等风险。

7.5 伦理审查

组建伦理审查小组，依据不伤害、有利、尊重、公正等医学伦理基本原则，评估衍生数据加工与使用的伦理合规性，重点审查是否侵犯数据主体权益、是否存在歧视或伤害等潜在风险。其中，以下衍生数据必须通过伦理委员会审查：用于精神疾病预测的，需防范标签化歧视；包含未成年人、孕产妇等特殊人群数据的，需强化权益保护；可能引发种族或性别歧视的分析结果类，需严格界定应用边界。

7.6 认定结果确认与登记

经过上述流程审核通过后，对符合认定标准的衍生数据予以确认，并进行统一登记，记录衍生数据的名称、类别、来源、加工方法、认定时间、适用范围等信息，纳入衍生数据管理体系，为其后续的使用、共享及监管提供依据；对未通过审核的，需明确原因并要求相关单位进行整改后重新申请认定。

附 录 A
（规范性）
医疗健康衍生数据认定申请表

A.1 基础信息

字段名称	填写要求	内容示例
申请单位	全称（与公章一致）+统一社会信用代码	北京 XX 医院（91110108XXXXXX）
数据联系人	姓名+职务+联系方式	张三（数据科主任）010-XXXXXXX
衍生数据名称	体现数据内容及类型的规范命名	"基于多中心 CT 影像的肺癌早筛 AI 模型训练数据集 V2.1"
应用场景	勾选：□临床诊疗 □医学研究 □公共卫生 □健康管理 □其他_____	□其他_新药临床试验筛选
数据量级	记录数/容量+时间范围	50 万例（2020.01-2023.12）

A.2 原始数据溯源

字段	填写说明	证明材料清单
数据来源	列明所有原始数据提供来源	1. XX 医院电子病历系统 2. XX 医院 PACS 系统 3. XX 智能手环监测数据
获取方式	勾选：☐患者授权 ☐机构合作 ☐公共数据库 ☐其他_____	☐患者授权（表格后附样本授权书）
原始数据类型	对应标准第 5 章分类	电子病历（5.1.1） 医学影像（5.1.2）
质量证明	提供数据质量评估报告编号	《XX 数据集质量评估报告》（QR-2024-XX

A.3 加工过程说明

1. 技术路线图

（需提供流程图，包含以下要素）：预处理方法、分析模型、验证方式

2. 关键参数记录表**（示例）**

环节	工具/算法	版本号	核心参数	输出校验值
数据清洗	OpenRefine	3.7.4	（1）去重规则：基于患者 ID + 检查编号联合去重 （2）缺失值填充：数值型字段（如年龄）用均值填充 （3）格式标准化：日期统一为 YYYY-MM-DD 格式，患者姓名去除首尾空格	SHA-256:a1b2c3.（实际需运行工具生成数据哈希，用于校验清洗后数据完整性 ）

环节	工具/算法	版本号	核心参数	输出校验值
特征提取	PyRadiomics	3.0.1	（1）voxelSize = 1×1×1mm （2）特征类型：形态学、一阶统计 （3）掩码处理：使用二值掩码限定 ROI 范围	MD5:d4e5f6...（记录特征矩阵的哈希值，验证提取结果一致性 ）

A.4 数据权属链声明页

一、声明事项

本单位确认，本申请表涉及的衍生数据生成过程符合以下权属要求：

1、原始数据权属：已通过 □患者知情同意书 □数据合作协议（协议编号） 获得合法使用权

2、加工过程授权：所有算法工具均已获得 □开源许可（GPL-3.0）□商业授权合同（附带 PDF 扫描件）

3、利益分配约定：衍生数据收益分配按 □机构协议 □患者授权条款 执行

二、权属凭证清单：

附件 1：患者授权书样本（含"允许用于衍生数据生成"条款）

附件 2：《XX 医院与 XX 公司数据合作框架协议》（盖章页）

附件 3：开源工具 LICENSE 文件截图

A.5 合规承诺

1、本数据 □包含 □不包含 个人信息,若包含则已按照 GB/T 35273-2020 完成去标识化（需附处理记录，处理记录见附表模板）

2、涉及人类遗传资源的，已通过科技部中国人类遗传资源管理办公室审批（批件号：_____）

3、承诺不将衍生数据用于□保险核保 □雇佣歧视 □其他_____等禁用场景

附表模板

去标识化处理记录（示例）

字段类型	原始字段名	处理方式	技术工具	重识别风险评估值
直接标识符	患者身份证号	哈希替换	ARX 3.9.0	0.05%

字段类型	原始字段名	处理方式	技术工具	重识别风险评估值
准标识符	就诊日期	泛化为季度	Python pandas	0.12%

附 录 B

（规范性）

典型衍生数据示例

B.1 影像组学特征数据集

属性	说明
原始数据来源	1.DICOM 格式医学影像（CT/MRI/PET-CT） 2.影像科 PACS 系统元数据
加工方法	1. 图像分割 2. 特征提取 3. 特征筛选
实质性改变	1. 内容改变：从像素值→定量影像特征 2. 形式改变：非结构化图像→结构化特征矩阵
示例输出	CSV 文件，含以下字段： 1.患者去标识化 ID 2.影像特征名 3.特征值
应用场景	1.肿瘤良恶性判别模型训练 2.放疗敏感性预测
合规要点	1. DICOM 头文件需删除 PatientID 等字段（符合 DICOM PS3.15 Annex E） 2. 特征提取过程需保留 ROI 标注记录

B.2 电子病历知识图谱

属性	说明
原始数据来源	1. 医院 HIS 系统结构化电子病历 2. 非结构化病程记录文本
加工方法	1. 实体识别（BERT-CRF 模型） 2. 关系抽取（规则引擎+深度学习） 3. 图谱构建（Neo4j 图数据库）
实质性改变	1.结构改变：离散病历→实体-关系网络 2.价值提升：实现疾病-症状-药品的语义推理
示例输出	图谱节点示例： 实体：[糖尿病，药物，二甲双胍] 关系：[糖尿病→首选治疗→二甲双胍，置信度=0.92]
应用场景	1.临床决策支持系统 2.药物不良反应挖掘
合规要点	1.需删除医生姓名等间接标识符 2.知识推理路径需经临床专家验证（附签字记录）

B.3 多中心临床研究数据集

属性	说明
原始数据来源	3 家三甲医院电子病历 中心实验室检测数据
加工方法	1. 数据标准化（CDISC SDTM 模型） 2. 跨源关联（主索引哈希匹配） 2. 差异校验（Kappa 检验）
实质性改变	1. 维度扩展：单一机构数据→多中心纵向数据 2. 质量提升：矛盾数据人工仲裁
应用场景	新药 III 期临床试验分析
合规要点	需提供各中心伦理批件且数据传输日志需保留

B.4 健康风险预测评分

属性	说明
原始数据来源	智能手环连续监测数据 体检报告
加工方法	1. 特征工程 2. 模型训练 3. 评分校准
实质性改变	1. 价值转化：原始体征→10 年心血管疾病风险概率 2.形态转化：时序数据→风险等级（A-E 级）
示例输出	JSON 格式： {"user_id": "MD5_3F2A...", "risk_score": 72, "grade": "C", "factors": ["夜间心率>80bpm", "血压波动>20%"]}
应用场景	1.健康管理 App 风险预警 2.商业健康保险核保
合规要点	1.需声明模型 AUC 等性能指标 2. 禁止用于就业歧视（需单独用户授权）

B.5 衍生数据特征对比表

类型	原始数据形态	衍生数据形态	价值跃迁点	典型合规风险
影像组学特征	DICOM 图像	定量特征矩阵	实现亚视觉信息挖掘	ROI 标注可能泄露病灶位置
电子病历知识图谱	分散病历文本	语义关联网络	支持临床推理	关系抽取错误导致医疗误读
基因注释库	原始测序数据	临床意义分级	转化碱基序列为诊疗依据	二次推断暴露家族遗传史
健康风险评分	连续监测波形	风险等级标签	从描述性数据到预测性输出	评分算法黑箱问题

使用说明：

1、示例选择原则：

覆盖影像、文本、生物信号、组学等多模态数据体现"实质性改变"的差异化实现路径

2、合规性标注要求：所有示例需在元数据中注明：

必填字段示例

```
derived_data_type = "影像组学特征"  # 对应本标准 6.1.4 条
compliance_evidence = ["GB/T 39725-2020", "WS/T 548-2017"]
```

3、动态更新机制：

每年补充新型衍生数据案例（如数字病理切片 AI 分析数据集）
案例库维护平台：国家健康医疗大数据中心（NHMDC）官网

参考文献

[1] GB/T 42439-2023 医疗数据安全指南
[2] GB/T 35273-2020 信息安全技术 个人信息安全规范
[3] 《中华人民共和国数据安全法》 [S]. 北京：中国法制出版社，2021.
[4] GB/T 25069-2022 信息技术 安全技术