健康医疗数据匿名化技术规范（试行）

为满足健康医疗数据高效合规流通的现实需求，在切实保障数据安全与个人隐私的前提下，需依托技术手段对数据进行匿名化处理，并落实相应的安全保障措施。为进一步规范匿名化处理活动，促进健康医疗数据的合理开发利用，特制定本规范。

1 适用范围

本规范适用于健康医疗数据开发利用过程中对个人信息的匿名化处理，可为健康医疗数据的匿名化工作提供指引，也可为监管部门进行数据开发利用监督管理提供参考。

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（包括所有的修改版）适用于本文件。

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 37964—2019 信息安全技术 个人信息去标识化指南

GB/T 39335—2020 信息安全技术 个人信息安全影响评估指南

GB/T 39725—2020 信息安全技术 健康医疗数据安全指南

GB/T 42460—2023 信息安全技术 个人信息去标识化效果评估指南

GB/T 32905—2016 信息安全技术 SM3 密码杂凑算法

GB/T 32907—2016 信息安全技术 SM4 分组密码算法

ISO 12052 医学影像数据成像与通信（DICOM）标准

3 术语和定义

下列术语和定义适用于本文件。

3.1

去标识化 de-identification

个人信息经过处理，使其在不借助额外信息的情况下无法识别特定自然人的过程。

注：去标识化处理后的信息也称去标识数据，其仍属于个人信息。

[来源：GB/T 35273—2017,定义 3.14]

3.2

匿名化 anonymization

个人信息经过处理无法识别特定自然人且不能复原的过程。

注 1：匿名化处理后的信息也称匿名数据，其不属于个人信息。

注 2：匿名化是去标识化的极端情况，强调无法识别特定自然人且不能复原。

3.3

个人健康医疗数据 personal health data

单独或者与其他信息结合后能够识别特定自然人或者反映特定自然人生理或心理健康的相关电子数据。

[来源：GB/T 39725—2020，定义 3.1]

3.4

健康医疗数据 health data

个人健康医疗数据以及由个人健康医疗数据加工处理之后得到的健康医疗相关电子数据。

[来源：GB/T 39725—2020，定义 3.2]

3.5

原始数据 raw data

初次产生或源头收集的、未经加工处理的数据。

[来源：国家数据局 数据领域常用用名词解释（第一批）]

3.6

匿名数据 anonymised data

经过匿名化处理后形成的数据。

3.7

直接标识符 direct identifier

数据记录中的属性，在特定环境下可以单独识别特定自然人。

[来源：GB/T 37964—2019，定义 3.7，有修改]

3.8

准标识符 quasi-identifier

数据记录中的属性，结合其他属性可唯一识别特定自然人。

[来源：GB/T 37964—2019，定义 3.8，有修改]

3.9

敏感属性 sensitive attribute

数据集中需要保护的属性，该属性值的泄露、修改、破坏或丢失会对个人产生损害。

[来源：GB/T 37964—2019，定义 3.10，有修改]

3.10

数据持有方 data holder

是指依法形成或合法获取健康医疗数据的权利人。

3.11

数据运营方 data operator

是指经授权或委托开展健康医疗数据运营的法人组织。数据运营方也可以是数据使用方。

3.12

数据使用方 data user

是指使用健康医疗数据产品和服务的法人组织或非法人组织。

3.13

完全公开共享 completely public sharing

数据一旦发布，很难召回，一般通过互联网直接公开发布。

[来源：GB/T 37964—2019，定义 3.12]

3.14

受控公开共享 controlled public sharing

通过数据使用协议对数据的使用进行约束。

注：例如通过协议禁止信息接收方发起对数据集中人体的重标识攻击，禁止信息接收方关联到外部数据集或信息，禁止信息接收方未经许可共享数据集。

[来源：GB/T 37964—2019，定义 3.13]

3.15

领地公开共享 enclave public sharing

在物理或虚拟的领地范围内共享，数据不能流出到领地范围外。

[来源：GB/T 37964—2019，定义 3.14]

4 总则

4.1 原则

健康医疗数据匿名化处理依据以下原则：

a）合法合规

遵守法律、行政法规等有关规定要求，尊重社会公德和伦理道德，不危害国家安全、公共利益、组织或自然人合法权益，保障自然人信息安全。

b）平衡效用

健康医疗数据专业性强，复杂度高，为充分实现数据要素价值，在满足安全合规的前提下，最大限度保留匿名化处理后数据的价值，力求实现隐私保护与数据开发利用价值之间的有效平衡。

c）分类分级

对于有条件开放或可开放的健康医疗数据，按照开发利用场景对匿名化处理程度、流通风险管理措施进行分类分级，促进不同场景下的数据合规流通。

4.2 业务要求

健康医疗数据持有方对原始数据进行治理，推动以患者为中心的全生命周期多模态数据关联（如同一患者多次就诊记录的关联或同一患者单次诊疗不同模态数据的关联）。根据不同的数据使用场景，选择适宜的技术对所需的原始数据进行匿名化处理，确保数据使用方无法识别特定自然人且不能复原。针对健康医疗数据流通利用的特定场景，采用保留必要字段（如性别）或对某些字段（如年龄、居住地）采取适度泛化等差异化技术手段，在保留数据价值的同时有效控制数据流通安全风险。

4.3 安全目标

健康医疗数据匿名化处理应实现以下安全目标：

a）无法识别

对原始数据无持有权的数据使用方无法提取特定自然人信息主体的记录，不能通过其已知的数据识别出特定自然人，不能将不同数据集中关于特定自然人的信息进行关联。

b）不可复原

对原始数据无持有权的数据使用方无法通过合理的手段还原出未经匿名化处理的原始数据。合理的手段，是指在当时技术条件下，合理可采用的识别手段，包括技术能力、经济成本、时间投入和获取信息的渠道等，仅在理论上存在可能的极端手段不纳入考虑。

c）风险可控

综合采取匿名化技术和流通管控措施防范健康医疗数据重识别、复原、滥用、泄露、篡改、

破坏、非法流通利用等风险，确保数据开发利用全过程可记录、可审计、可追溯，定期跟踪评估和持续改进。

4.4 相关主体

相关主体包括数据持有方、数据运营方和数据使用方：

a）数据持有方

履行安全与合规义务，做好多源数据的有效整合，确保原始数据的完整性与一致性。根据具体开发利用场景，选择恰当的匿名化技术，对数据进行匿名化处理，实现数据效用与隐私保护的平衡。匿名化处理后的数据不属于个人信息，可在遵守国家相关法律法规要求的前提下安全流通至数据使用方。

b）数据运营方

数据运营方应依法依规在授权或委托范围内开展数据资源开发、数据产品经营和技术服务，承担数据安全主体责任。在数据流通过程中，数据运营方应依据数据开发利用场景和合规约束，对数据进行相应匿名化处理。

c）数据使用方

依法合规对数据进行开发利用，确保开发利用环节的安全与合规责任。

4.5 个人健康医疗数据范围与分类

4.5.1 个人健康医疗数据范围

参照 GB/T 39725—2020，个人健康医疗数据包括但不限于：

a）提供健康医疗服务时登记的个人信息。

b）出于健康医疗目的，例如治疗、支付或保健护理等，分配给个人的唯一标识号码或符号等。

c）在向个人提供健康医疗服务过程中采集的有关个人的任何数据，例如既往病史、社会史、家族史、症状和生活方式等各类病历记载的数据。

d）来自身体部位或身体物质，例如组织、体液、血、尿、便、气体，以及 DNA、RNA、蛋白质等生物大分子、代谢小分子、肠道微生物等检查或检验的结果数据。

e）可穿戴设备采集的与个人健康相关的数据，并且该种数据：

1）本身或者明显为健康医疗相关数据；

2）或是由传感器采集的，可以单独或者与其他数据结合用来对可穿戴设备的用户的健康状况或者疾病风险进行判断的数据；

3）或是可穿戴设备采集的数据，为对用户的健康状况或者疾病风险进行判断后的结论；

4）或是通过可穿戴设备相连的 APP 或者系统进行提供的，并非可穿戴设备使用者另行提供的。

f）接受的健康医疗服务相关数据，例如检验检查、医嘱、诊断、操作、药物、医疗效果等。

g）为个人提供健康医疗服务的服务者身份数据。

h）关于个人的支付或医保相关数据。

i）医学科研个人相关数据，例如临床研究病例数据、生物样本库、全基因组等多种生物组学测序结果、医学相关队列研究结果等。

j）公共卫生与预防医学数据，例如疾控中心、公共卫生管理部门收集的疾病卫生监测个人数据。

k）妇幼保健数据，例如妇幼保健院、医疗卫生机构等收集的妇幼保健服务与健康管理数据。

4.5.2 个人健康医疗数据可识别程度分类

个人健康医疗数据根据可能识别特定自然人的程度进行分类，包括：

a）含有直接标识符的；

b）含有准标识符的；

c）不属于 a）和 b），但可能包含特定自然人其他敏感属性的。

5 匿名化实施流程

5.1 匿名化实施流程概述

健康医疗数据匿名化实施流程如下：

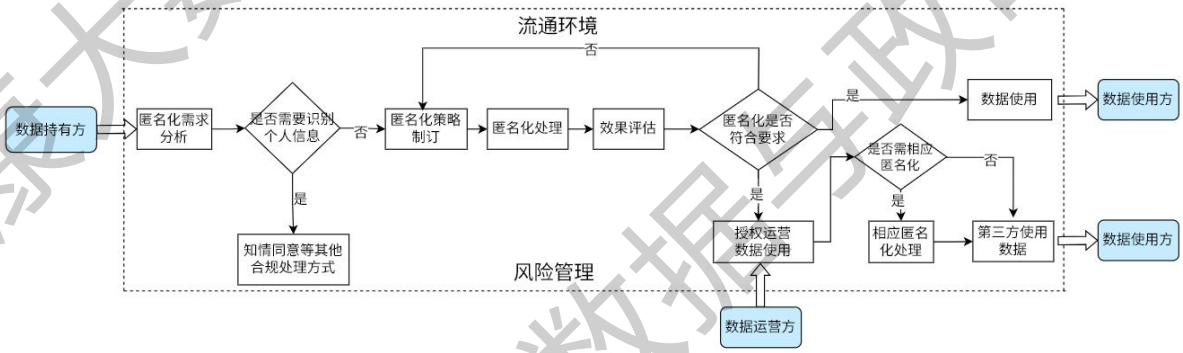

图 1 匿名化实施流程

5.2 匿名化需求分析

数据持有方根据数据开发利用场景，分析业务需求的数据范围、使用对象、使用方式等，判断匿名化处理的必要性和可行性。根据业务场景对数据安全性和可用性的要求设定匿名化目标，明确各方可接受的最低风险程度和满足后续用途的最低要求，并根据内、外部的环境变化动态调整。若数据使用需要识别个人信息，须通过取得个人知情同意等方式获得处理个人信息的合法性基础。

5.3 匿名化策略制订

在匿名化需求分析基础上，根据应用场景，分析个人信息中的直接标识符、准标识符以及数据集中的其他敏感信息，制订应用场景所需的匿名化处理策略，选择合适的技术（例如删除、泛化、随机化）和模型（例如 k-匿名、差分隐私），并确定相应的参数，参数确定参见附录 A。明确数据安全技术措施，防范来自场景内外的安全风险，并在实施前对策略进行审核评估。

5.4 匿名化处理

按照匿名化处理策略，选择相应匿名化技术开展处理转换，涉及需全生命周期多模态数据关联需求的，应在匿名化技术处理之前做关联。匿名化技术措施可参考 GB/T 37964 的相关要求，包括删除、统计、加密、假名化、泛化、随机化、数据合成等技术，不同技术方法也可组合使用，

具体处理方法和技术参见 6.1 和 6.2。

5.5 匿名化效果评估

对数据匿名化从“无法识别”和“不能复原”两个维度进行效果评估，基于特定业务场景和数据处理环境，对匿名化结果进行评估，确认输出数据是否属于匿名数据，达到了有效保护个体隐私的目的，并保留数据可用性。并对评估过程和结果进行记录，形成匿名化评估报告。若匿名化效果不达标，则重新制定匿名化策略和匿名化处理。匿名化效果评估参考附录 A。

随着后续的使用以及技术的发展，原先匿名化的数据如果被证明可以重新识别到特定自然人，处理者还应当持续评估匿名化处理后的个人信息重识别风险。

5.6 匿名化风险管理

数据持有方、数据运营方、数据使用方应做好数据匿名化风险管理。采取合同协议、日志记录、合规审计、风险监测等措施，对数据利用安全风险进行持续管理，将数据安全风险控制在可接受风险水平。

5.7 数据流通环境

数据流通环境为数据匿名化处理、加工利用的安全可信环境，应采用相应的技术和管理措施，确保数据流通的安全。

6 匿名化处理方法

6.1 常用匿名化处理技术

6.1.1 抽样

描述：通过选取数据集中有代表性的子集对原始数据集进行分析和评估。对数据集进行随机抽样能够增加识别出特定个人信息主体的不确定性，从而提高后续应用其他匿名化技术的有效性。

适用场景：从大样本的数据集中抽取部分数据进行匿名化处理。

注意事项：从数据集中抽取样本的方法很多，各方法差异很大，需根据数据集的特点和预期的使用场景来选择，以保证样本对原始数据集的代表性。

示例：从 1000 万份记录中随机抽取 1 万份记录作为处理对象，在保留代表性的同时大幅降低数据处理成本。

6.1.2 记录删除

描述：从数据集中删除或置空不满足匿名化要求的记录或数据项。

适用场景：少量记录因为异常数值无法满足匿名化标准。

注意事项：可能对数据集的统计属性（如平均值、中位数、方差等）造成影响。

示例：在研究队列中删除年龄为稀有值（如 100 岁以上）的记录。

6.1.3 属性删除

描述：删除或置空整个属性（列），通常是标识符或不必要字段。

适用场景：删除对数据分析无用但可能泄露隐私的标识符或不必要字段（如姓名、身份证号、职业、教育水平）。

注意事项：应确保数据源在导出时就不包含不必要的属性。

示例：在研究队列中删除研究对象姓名、身份证号、职业、工作单位等与研究无关的字段。

6.1.4 随机化

描述：通过随机修改属性的值，使得随机化处理后的值区别于原来的真实值。

适用场景：当属性不允许删除或置空时，可以采用随机化技术给属性赋值，从而消除属性本身的意义。

注意事项：尽管随机化消除了数据的真实性，但仍需选择合适的随机化参数避免不同数据随机化后出现相同数值。

示例：当患者姓名字段不允许为空的情况下，可以采用随机化方式给姓名赋值代替原有数值。

6.1.5 字符遮盖

描述：用特定符号（如“*”或“#”）替换部分字符，遮掩敏感信息，如“123456”变为“123**”。

适用场景：当属性的部分信息足以满足用途，且遮掩部分能提供所需的匿名化程度。常用于展示地址等敏感字段。

注意事项：需要考虑原始数据的长度是否会泄露信息。

示例：将地址“北京市朝阳区十里堡甲 X 号院 X 栋 XXX”遮盖为“北京市朝阳区******”。

6.1.6 假名化

描述：用生成的假名替换标识符，可以采用分配表建立映射关系，也可以采用可逆加密和不可逆加密技术，不可逆的假名化安全程度更高。

适用场景：需要保持数据记录间的关联性或者可追加性（如患者多次就诊记录的关联或更新），但无需识别真实身份。

注意事项：需安全存储映射关系，防止泄露导致隐私风险。

示例：将姓名“张三”、“李四”假名化为“ABC”、“123”。

a）可逆加密

描述：利用算法和密钥将标识符加密为假名，在获取算法和密钥后可以将假名还原为标识符。要求采用国密算法。

适用场景：需要对单条记录利用算法和密钥还原标识符。

注意事项：如果多组数据采用相同的算法和密钥，会导致较大的隐私风险。

示例：利用可逆加密将姓名“张三”、“李四”假名化为“ABC”、“123”。获取算法和密钥后，可以将单条假名“ABC”还原为“张三”。

b）不可逆加密

描述：通过特定数学变换将标识符转化为无法还原的密文，加密后的数据无法通过逆向计算获取原始信息，仅能通过相同算法比对验证结果有效性。

适用场景：不需要对单条记录利用算法和密钥还原标识符。

注意事项：仅能实现标识符到假名的单向映射，关联效率低。

示例：利用不可逆加密将姓名“张三”“李四”假名化为“DEF”“456”。无法将假名“DEF”解密为标识符“张三”。只有重新将数据集采用相同算法单向加密，对比加密后的假名才能回溯“DEF”对应的原始标识符为“张三”。

6.1.7 泛化

描述：降低数据值的精度，将具体值替换为范围或类别。

适用场景：数据精确值不是分析的关键因素，泛化后仍能满足用途的场景。常用于时间（如具体日期替换为年份）、地点（如具体地址替换为城市）或数值数据（如年龄）。

注意事项：需合理选择泛化范围，避免过度泛化导致数据失去实用性。泛化层级过低可能无法有效保护隐私，过高则可能削弱数据分析效果。

示例：将“95 岁”泛化为“80 岁以上”。

6.1.8 扰动

描述：通过向数据值添加微小随机变化，增加隐私保护。

适用场景：对数据的精确值要求不高，允许一定程度的误差。适用于分析中对数据分布更关注的场景，如聚类或回归模型。

注意事项：扰动幅度需与数据属性的范围相适应，过大可能失真，过小则保护不足。可能影响某些分析模型的精度，需权衡隐私与分析质量。

示例：将患者的入院日期、出院日期随机偏移 2-3 天，在住院时长不变的情况下保护患者隐私。

6.1.9 聚合

描述：将多个记录的数据汇总为统计值，如总数、平均值或中位数，并用统计值代替原有数值。

适用场景：数据分析仅依赖统计值，个体记录的具体内容不影响整体分析。

注意事项：数据聚合可能会降低数据的有用性。因得到的数据是统计值，无法反映独立数据记录的特征。

示例：当体重对整体数据分析影响不大时，可以用人群体重的平均值代替每个人的具体体重值。

6.2 特殊匿名化处理技术

6.2.1 DICOM 标签数据处理

描述：使用符合 DICOM 标准的去标识化工具，参考 DICOM PS3.15 Annex E 定义的“基本去标识化配置文件”（Basic Application Level Confidentiality Profile）执行。在保留医学影像诊断价值的前提下，对可识别患者身份的标签数据进行匿名化处理。

适用场景：数据分析需要使用 DICOM 标签中的数据。

注意事项：

a）DICOM 标签为文本，匿名化处理方式可以参考文本的匿名技术。

b）应默认清除所有私有标签（Private Tags），除非经安全评估确认无隐私风险。

c）匿名化后应生成新的 Study Instance UID(0020,000D)、Series Instance UID(0020,000E)、SOP Instance UID(0008,0018)，避免与原始影像关联；

d）推荐在匿名化流程末尾执行 DICOM 一致性校验，确保文件仍可被标准 PACS 读取。

示例：假名化 DICOM 标签中的 Patient ID、Patient Name、Institution Name 等数据。

6.2.2 检查影像/图像/视频抹除与遮蔽

描述：通过在图像/视频中添加噪声、遮挡物或模糊区域，实现图像中的敏感信息不可识别。

适用场景：图像/视频中包含个人信息标识符或敏感数据的情形。

注意事项：应确保数据源在导出时，图像/视频中不包含个人信息标识符。

示例：对可能包含面部特征的影像（如头颅 MRI、头颅 CT、牙科全景片），采用面部区严禁域自动检测与模糊/遮蔽技术。

6.3 按数据可识别程度分类实施的匿名化技术

个人健康医疗数据常用直接标识符匿名化技术见表 1，个人健康医疗数据常用准标识符匿名化技术见表 2，个人健康医疗敏感数据匿名化技术见表 3。

表 1 个人健康医疗数据常用直接标识符匿名化技术

编号	常见直接标识符	描述	建议匿名化技术
1	姓名	本人及相关人员姓名，包括患者姓名、家属姓名、医生姓名等	属性删除/不可逆加密
2	证件号码	可以识别个人及相关人员身份的证件号码，包括身份证号、护照号、军官证号、驾照号、居住证号、工作学习编号、车牌号、车辆识别号、社会保障卡号、残疾证号等	属性删除/不可逆加密
3	医疗身份号码	医疗行为中用于识别患者身份的号码，包括健康卡号、就诊流水号、挂号编号、门诊编号、住院编号、病案号、健康档案号等	属性删除/不可逆加密
4	联系方式	可以联系个人的通讯号码，包括电话号、传真号、电子邮件、网络账号及昵称	属性删除/字符遮盖
5	详细住址	可以精准定位到个人居住地的详细住址信息	属性删除/泛化/字符遮盖，保留省、市、区（县）、乡镇(街道)部分
6	生物识别码	可用于识别个人身份的生物特征，包括指纹、声纹、掌纹、耳廓、虹膜、面部特征等	属性删除/特殊处理
7	其他	包括支付账号、个人设备标识符、互联网协议地址号（IP）、网络通用资源定位符（URL）等	属性删除

表 2 个人健康医疗数据常用准标识符匿名化技术

编号	常见准标识符	描述	建议匿名化技术
1	健康指标	个人健康数据中相对稳定的指标，包括性别、年龄、身高、体重、血型、肺活量等	属性删除/泛化/聚合
2	健康事件时间	健康医疗相关事件发生的时间，包括门诊、住院、手术、检查、检验、用药、治疗、出院、随访、死亡等事件发生时间	属性删除/扰动/泛化
3	健康事件地点	健康医疗相关事件发生的地点，包括机构、场所、地理位置等	属性删除/字符遮盖/泛化
4	单位信息	工作或学习单位，包括工作单位、学校、班级等	属性删除/假名化
5	其他个人属性	包括国籍、籍贯、婚姻状况、受教育水平、职业、户籍等	属性删除/假名化

表 3 个人健康医疗敏感数据匿名化技术

编号	敏感数据	描述	建议匿名化技术
1	标识符以外的健康医疗数据	个人健康医疗数据集涵盖门诊/住院病历、诊断结论、手术记录、用药处方、检验检查报告、疾病编码（如	数据如包含直接标识符、准标识符按照相应匿名化处

编号	敏感数据	描述	建议匿名化技术
		ICD）、治疗计划等临床诊疗相关内容数据	理
2	DICOM 影像	指符合医学数字成像与通信（Digital Imaging and Communications in Medicine, DICOM）标准的医学图像及其元数据，包括但不限于 CT、MRI、X 光、超声、PET 等模态的原始影像文件。此类数据通常包含患者标识信息（如 Patient ID、Patient Name）、检查时间、设备参数、医疗机构信息等嵌入式元数据，部分图像像素本身也可能隐含可识别特征，具有较高的隐私泄露风险	图像/视频抹除与遮蔽，DICOM 标签数据处理
3	医学图片/视频	非 DICOM 格式的其他各类医学图像、视频等非文本数据	图像/视频抹除与遮蔽

注：人类遗传资源信息匿名化遵从《中华人民共和国人类遗传资源管理条例》及相关法律法规要求处理。

7 匿名化效果评估

数据匿名化效果评估的目标在于确保数据在去除或隐藏个人身份信息后，能够有效保护个体隐私，并保留数据可用性。

7.1 评估维度

匿名化处理效果从“无法识别”和“不能复原”两个维度进行综合评估。

a）无法识别：评估处理后的数据集，在综合考虑数据使用场景和使用环境安全管控措施的情况下，是否满足数据使用方无法识别原始数据所描述的特定自然人的要求。

b）不能复原：评估处理后的数据集或在所属环境下输出的计算结果，综合考量现存的、公开的、可预期的方法和条件，数据使用方是否有能力通过逆向回溯的方式，实质性恢复成数据处理前的状态，可结合技术难度、实施成本、可能风险等方面，论证关键标识符或属性是否具有实质性复原的合理可能。

7.2 评估方法

匿名化处理效果可采用以下一种或多种方法进行评估：

a）标识符识别法：参考 GB/T 37964—2019 规定的方法，通过查表识别法、规则判定法、人工分析法，判断数据集中是否包含目标标识符。

b）匿名化度量参数计算法：

1）K-匿名值：关注匿名数据集中每个等价类存在相似记录的数量，K-匿名值越高意味着重标识风险越低，K-匿名值越低意味着风险越高。基于 K-匿名值进行评估时，宜结合场景系数和环境系数来判断。

2）L-多样性：在 K-匿名的基础上，要求每个准标识符分组内的敏感属性（如疾病类型）至少有 L 种不同值，防止敏感信息被推测；

3）T-接近性：核心思想是使每个组（等价类）中敏感记录的分布与原始数据集中相应的分布足够接近。具体来说，根据 T-接近性原则，原始数据集中某个属性的分布与组内同一属性的分布之间的距离应小于或等于 T。

c）模拟攻击测试法：根据可能面临的实际威胁场景，模拟外部入侵者和内部违规人员尝试重新识别匿名数据的行为，判断是否能够通过攻击进行重识别。

d）技术成本分析法：在合理可能的技术条件和资源情况下，分析重识别的可能性，包括现存的、公开的、可预期的技术手段能否进行重识别或存在相关威胁，以及具体场景下重识别所需的

经济成本、时间成本等是否合理；

e）专家决策法：通过邀请法律、数据安全合规、技术研发等领域的专业机构及权威专家，基于外部独立且专业的经验和知识判断，对匿名化处理后的效果进行评价并提供相关证明。

7.3 实施效果评估

基于特定业务场景和数据处理环境，根据 7.2 中评估方法，对匿名化实施效果在 7.1 中各个维度进行评估。

7.3.1 个人信息标识度效果分级

基于数据是否能直接识别个人信息主体，或能以多大概率识别个人信息主体，个人信息标识度分级划分为 4 级，详见表 4，用于区分个人信息去标识化效果。

表 4 个人信息标识度 4 级划分

分级	划分依据
1 级	包含直接标识符,在特定环境下能直接识别个人信息主体
2 级	消除了直接标识符,但包含准标识符,且重标识风险高于或等于可接受风险阈值
3 级	消除了直接标识符,但包含准标识符,且重标识风险低于可接受风险阈值
4 级	不包含任何标识符

7.3.2 匿名化评估流程

对个人信息进行匿名化处理后，实施匿名化评估的流程，流程图如图 2。实施过程包括：

a）进行去标识化效果评估。

b）进行对抗性测试和不能复原性核验。

c） 对于未通过对抗性测试和不能复原性核验的数据集，需要重新进行匿名化处理及后续步骤。

d）通过对抗性测试和不能复原性核验后，出具匿名化评价报告。

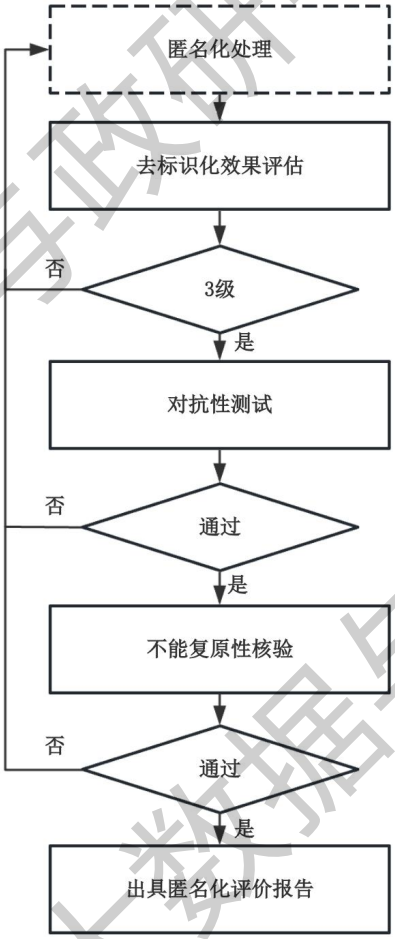

图 2 匿名化评估流程

7.3.2.1 去标识化效果评估

参照 GB/T 42460—2023，对处理后的结果数据进行去标识化效果评估，如未达到 3 级（即消除了直接标识符，但包含准标识符，且重标识风险低于可接受风险阈值），则调整去标识化技术或模型，继续对其进行去标识化处理，直至获得的结果数据集个人信息标识度达到 3 级。

7.3.2.2 对抗性测试

对抗性测试是模拟潜在攻击者通过各种手段尝试重新识别匿名化数据中的个人信息，以验证数据集在实际攻击下的安全性。对于匿名化处理后个人信息标识度达到3级的结果数据集，基于所限定的使用场景，按如下步骤进行对抗性测试。

a）定义模拟攻击者：根据可能面临的实际威胁场景，定义模拟攻击者的来源（内部、外部）、动机、技能、资源、目标和合理可能使用的任何攻击手段。模拟攻击者需建立在预定的数据应用场景，包括目的、范围、环境约束等；

b）识别关键变量：分析结果数据集中除直接标识符以外的各个属性列的数据属性标识度，对超过一定阈值的属性，重点分析其是否可作为用于攻击测试的关键变量；

c）攻击测试准备：根据攻击者模型，收集攻击者可能拥有的任何可以合法获得的数据，准备攻击测试所需的数据和环境。合法获得的数据资源包括开放的数据资源和私有数据资源；

d）实施攻击测试：模拟攻击者，利用可用数据和技术，对结果数据集进行识别和关联等攻击；

e）根据攻击测试的结果，评估结果数据集的安全性，如果在测试中发现能够通过攻击在结果数据集中使得重标识风险大于设定阈值，则测试不通过，需要重新调整去标识化策略和方法。

7.3.2.3 不能复原性核验

不能复原性核验的目的在于证明匿名化处理后的结果数据在合理可得的技术与资源条件下不能被恢复为原始个人信息。核验侧重于技术不可逆性、关键辅助材料（如密钥、对照表等）不可得性与环境不可越权三个方面的证据性审查。核验范围应覆盖生产、备份与灾备环境及可访问副本。核验实施包括核验准备、技术不可逆性核验、关键辅助材料不可得性核验、环境不可越权核验。判定规则如下：

a）同时满足下列条件时，判定为通过：

1）处理方法层面不存在可逆映射或已执行不可逆收口；

2）关键辅助材料不可得（已销毁或隔离到等同不可得的强度），并有证据；

3）环境隔离、访问控制与日志审计有效且证据完备。

b）出现下列任一情形，判定为不通过：

1）发现可逆路径或可获取之关键辅助材料；

2）证据缺失或无法复现；

3）备份/快照/日志中存在足以复原的材料且未整改。

c）再评估与复验触发，当出现下列情形之一时，应开展再评估与复验：

1）处理方法或关键参数变更；

2）新引入或撤销关键辅助材料；

3）接收方环境风险变化或数据用途变化；

4）发生疑似重标识或泄露事件。

7.3.2.4 出具匿名化评价报告

匿名化处理和效果评估应实行双人复核机制，数据匿名化处理责任人与评价/批准责任人相互独立。评价/批准责任人不得参与匿名化处理实施及关键辅助材料的日常保管。必要时可委托内部合规/审计或第三方出具独立复核意见。

匿名化评价作为正式验收/背书的依据，由独立角色（内部合规/第三方）完成，包括无法识别的评价、不能复原的评价、对抗性测试评价、综合评价等评价，并出具匿名化评价结论，评价报告可引用匿名化处理及评估中各阶段的评价记录与证据材料。

8 匿名化保障措施

8.1 数据流通环境技术措施

应通过技术手段构建安全可信流通环境，确保匿名数据合规流通。

a）基础与安全计算

1)基础安全设施：应构建满足网络安全等级保护三级要求及相关行业标准的基础防护体系，具备网络隔离、数据防泄露及抗攻击能力。

2）安全计算环境：应根据业务场景的敏感度，部署支持多方安全计算、可信执行环境或联邦学习等隐私计算技术，确保数据在计算过程中不明文暴露。对不同来源的数据进行存储隔离，确保数据独立存储、权限独立管控。

b）访问控制与身份认证

1）强身份鉴别：采用多因子认证或数字证书技术，具备严格的身份审核与接入校验措施。

2）微隔离与最小授权：设置基于角色或属性的访问控制策略，通过计算沙箱、虚拟专用网等技术实现租户间的逻辑隔离，确保访问权限精确至最小必要范围。

c）安全审核与过程监控

1）计算逻辑审核：部署第三方加工软件（含脚本、算法模型等）的静态代码审计与动态沙箱监测设施，识别并拦截可能导致数据隐蔽传输或重识别的恶意逻辑。

2）全链路审计：系统应自动记录环境内所有主体的数据调取、模型训练及参数导出行为。审计日志应具备抗篡改性，并对敏感操作执行实时预警。

d）匿名化验证与输出管控

数据输出前，环境应内置匿名化效果评估工具，确保输出结果符合匿名化合规要求。

e）应急响应技术支撑

1）态势感知：部署安全态势感知系统，实现安全事件的自动发现、响应与处置。

2）数据追溯：利用水印技术或区块链存证，确保存储于环境内及流出的数据具备可追溯性，为事件调查提供支撑。

8.2 匿名化管理措施

建立匿名化管理机制，明确匿名数据流通相关方的职责，促进匿名化工作高效开展。

8.2.1 数据持有方管理

数据持有方管理要求包括：

a）应建立数据流通管理制度并定期更新，明确数据匿名化、数据防泄漏等安全保障措施；

b）应对数据使用方的使用场景、目的和处理过程进行充分了解和审核，确保数据使用申请满足数据使用场景所需的最小范围数据；

c）应组织具备匿名化专业技术能力的团队开展数据匿名化，明确数据匿名化处理的人员和职责，定期对负责人员进行相关技术培训；

d）应对数据使用场景和数据匿名化处理的合规性和安全性进行评估，确保提供的数据符合法律相关要求；

e）应采用合同协议等具有约束效力的形式，明确数据流通相关方的责任和义务，包括但不限于：

1）数据使用场景、使用目的和范围限制；

2）数据安全保护义务和责任；

3）禁止数据使用方对匿名数据进行重标识或复原；

4）明确是否允许数据使用方将数据进行二次提供及其条件；

5）数据泄露事件的通知和应急响应机制；

6）违反数据流通利用约定的处罚措施。

f）应制定并定期演练数据匿名化流通应急预案，发生数据复原、泄露、滥用等安全事件时，应立即启动预案，采取措施防止危害扩大，消除安全隐患，并按照规定向有关部门报告。

8.2.2 数据使用方管理

数据使用方管理要求包括：

a）应按照最少够用和目的限制原则，申请满足实际使用目的的最小必要数据，限制匿名数据的访问权限；

b）应与数据持有方等签订数据使用合同协议，合同协议应满足 8.2.1 中 e）要求；

c）应采取技术措施禁止任何试图重新识别特定自然人、复原原始数据的行为，并确保采取适当措施销毁任何意外重新识别的个人信息；严禁将个人信息隐藏在输出结果中。

d）应对接触匿名数据的人员开展培训，建立离岗/变更权限回收机制；并对匿名数据的开发利用操作进行监管，禁止超范围使用或滥用数据等行为；

e）达成使用目的或期限届满后，应删除匿名数据，并确保不可恢复。

8.2.3 数据运营方管理

数据运营方管理要求包括：

a）在开展数据资源开发和技术服务时，应遵循 8.2.2 管理要求；

b）应提供满足 8.1 要求的安全可信流通环境；

c）应向数据流通利用相关方告知数据流通环境提供的安全技术能力和安全管理措施；

d）应定期对所提供的流通环境技术措施进行安全评估，确保其提供的安全保障能力满足数据匿名化流通安全需求，不会造成数据泄露、未授权访问、重识别、复原等；

e）应按照最小授权原则对数据流通环境设置严格的访问控制策略，并确保访问控制策略的实施；

f）应对数据持有方、数据使用方进行身份审核和鉴别；

g）应建立第三方加工软件监测机制，对其在流通环境中部署和运行等情况进行审核，防止数据泄露、滥用；

h）应对匿名数据的开发利用操作进行日志记录，并定期对操作行为进行安全审计；

i）应在离开安全可信环境前对输出数据进行匿名化效果评估；

j）应制定并定期演练数据匿名化流通应急预案，出现安全事件时及时响应，采取措施消除安全隐患并防止危害扩大，对违规相关方采取暂停交易、下架产品、列入黑名单等措施，并向数据持有方及有关部门报告。

北京市卫生健康大数据与政策研究中心

2025 年 12 月 31 日

附 录 A
（资料性）
基于K匿名的效果评估方法

A.1 评价方式

对于离线库表结构的数据集，建议基于K匿名进行效果评估。具体计算方式见公式A.1：

$$A = K \times S \times E \quad \text{(A.1)}$$

其中：

A——匿名化程度，表示数据集的匿名化程度。其值大于等于1时，认为满足匿名化要求；

K——数据集K匿名值，表示数据集经过匿名化处理后，具备相同的准标识符字段组合的记录的条数的最小值；

S——场景系数，表示数据匿名化后使用场景的安全系数，如领地公开共享、受控公开共享、完全公开共享等；

E——环境系数，表示数据流通时，数据流通环境的技术保障能力和管理保障能力。

A.2 场景系数

通常而言，应结合数据公开共享类型确定场景系数，例如完全公开共享应设置尽可能高的K匿名值；受控公开共享的K匿名值可以小于完全公开共享的标准；而领地公开共享的K匿名值可以相对较低。相关场景系数建议见表A.1。

表A.1 场景系数建议

流通范围	场景	建议的场景系数
领地公开共享	组织内部同一个事业群的数据流通	1/3
领地公开共享	组织内部跨事业群的数据流通	1/4
受控公开共享	组织外部两方的数据流通	1/5
受控公开共享	组织外部多方的数据流通	1/6
完全公开共享	对外公开	1/20

注：在实际使用时，可根据具体的使用场景，适当调整。

A.3 环境系数

对于环境保障能力，宜采用定性评估的方式。根据环境保障能力确定环境系数，环境系数的中间值建议为1。环境保障能力的评估维度包括技术保障能力和管理保障能力。

a）技术保障能力：包括安全和隐私控制能力，如，是否采用权限管理、访问控制策略等；数据流通技术保障能力，如是否采用安全隔离、数据沙箱、封闭域、专区、隐私计算等技术；抗击重识别攻击的能力，如是否采用技术手段抗重识别攻击等；

b）管理保障能力：包括组织建设、制度流程、人员能力、协议合同约束、审计机制、事件与应急管理、风险控制能力等管理保障能力的综合评估。

附 录 B
（资料性）

个人健康医疗数据匿名化参考案例

案例：向人工智能科技公司提供精标注的专病高质量数据集用于模型训练场景

B.1 场景描述

某医院作为国家结核病临床重点专科单位，已系统性积累并高质量标注了约 2 万例肺结核的胸部 CT 影像数据，每例均包含 DICOM 原始图像、病灶边界、空洞/结节/磨玻璃影等结构化标注结果。为推动人工智能在传染病早筛领域的应用，医院拟将该数据集向具备医疗 AI 研发资质的影像科技企业进行合规流通，用于训练和验证肺结核智能筛查与辅助诊断模型。

a）场景相关方：包括数据持有方、数据使用方。

b）涉及的数据类型：病灶标注结果的 CT 影像数据，包括患者信息、检查信息、序列信息、图像信息、设备信息。

B.2 可行性分析

本场景数据匿名化处理具有可行性。首先，涉及的数据主要为静态的胸部 CT 影像数据，不涉及动态更新内容和颈部以上的 CT 扫描数据，不存在脸部及头部（数据重建）特征识别的风险，技术实现难度相对可控；其次，临床数据中与研究最相关的医学特征（如检查影像等），可以在匿名化过程中保留，不会严重影响数据可用性；再次，数据流通范围仅限于模型训练企业内部，使用环境可控，降低了复原风险；此外，所需计算资源和技术成本在合理范围内，不会对项目造成过度负担。

B.3 处理对象

B.3.1 数据范围

本场景数据匿名化处理范围主要涵盖确诊肺结核患者的标注后 CT 影像数据。一方面，这些数据直接支持肺结核模型智能筛查、诊断技术研究这一核心业务需求；另一方面，这些数据中包含了可能导致患者被识别的多种信息，需要进行匿名化处理以确保合规。具体数据类型包括：

a）患者信息数据：患者姓名、患者 ID、患者性别、患者年龄等。

b）检查信息数据：医疗机构名称、患者出生时间、患者出生日期、检查号、检查日期、检查实例号、扫描部位等。

c）序列信息数据：序列实例号、序列日期等。

B.3.2 数据性质

结合数据集的使用场景，对肺结核患者临床数据集中的直接标识符、准标识符和敏感属性进行识别，具体表 B.1 所示：

表 B.1 肺结核患者个人健康医疗数据标识符示例

数据属性	数据字段
直接标识符 （DICOM 标签）	患者姓名
	患者 ID
	检查号

数据属性	数据字段
	检查实例号
	序列实例号
准标识符 （DICOM 标签）	医疗机构名称
	患者性别
	患者年龄
	患者出生时间
	患者出生日期
	检查日期
	扫描部位
	序列日期

B.3.3 处理目标

为保障临床数据的最低风险且满足后续模型训练的最低使用要求，匿名化处理目标的设定考量包括以下方面：

a）为保证匿名化处理后的数据符合安全性要求，根据数据流通场景、流通环境及数据敏感程度（医疗敏感数据）的综合评估，将风险控制目标设定如下：匿名化程度 A 值>1，基于数据集样本量（2 万例）、准标识符组合种类、数据流通范围以及流通环境的技术保障能力和管理保障能力分析确定；

b）为保证匿名化处理后的数据可满足后续模型训练的最低要求，将数据的可用性目标设定如下：

1）保留性别、年龄段等与疾病相关的人口学特征；

2）保留时序性检查过程的分析价值；

3）确保数据可用于模型训练。

c）基于上述目标，对 DICOM 数据中相关标识符的匿名化处理方法如下：

1）直接标识符：全部删除或替换为不可逆的编码；

2）准标识符：医疗机构名称、患者出生时间、患者出生日期、扫描部位，进行属性删除或替换为不可逆的编码；患者年龄进行泛化处理；检查日期、序列日期进行扰动处理；患者性别，保留原值。

B.4 匿名化准备

B.4.1 策略制定

针对不同类型的数据采取差异化处理。其中，对直接标识符进行属性删除或假名化处理；对准标识符进行泛化或扰动处理；对像素数据进行随机化处理。

B.4.2 处理技术选择

a）属性删除：对患者姓名等直接标识符进行置空；对医疗机构名称、患者出生时间、患者出生日期、扫描部位等间接标识符进行置空；

b）假名化：对患者 ID、检查号、检查实例号、序列实例号进行不可逆的假名化处理，替换为伪码，保证患者具有唯一 ID；

c）泛化：对患者年龄进行泛化处理，将“患者年龄”从“53 岁”泛化为“5X 岁”；

d）扰动：对检查日期、序列日期进行扰动处理，将“检查日期”和“序列日期”向后偏移特

定天数；

e）随机化：对某些像素数据采取随机化处理。

B.5 分类处理

CT 影像数据匿名化处理措施如下：

a）直接标识符：Patient Name（患者姓名）采用属性删除处理；Patient ID（患者 ID）采用假名化处理，假名后的形式为“P_4 位数字_5 位数字”；Accession Number（检查号）、Study Instance UID（检查实例号）、Series Instance UID（序列实例号）采取假名化处理；

b）准标识符：Institution Name（医疗机构名称）、Patient Birth Date（患者出生日期）、Patient Birth Time（患者出生时间）、Body Part Examined（扫描部位）采用属性删除处理；Patient Age（患者年龄）、Study Date（检查日期）、Series Date（序列日期）采用泛化处理；

c）其它数据：设备信息、像素数据不属于个人信息，但仍对一部分此类数据采用属性删除、随机化及假名化处理。

B.6 环境保障

为控制匿名化数据后续使用过程中可能产生的风险，采取相应技术和管理措施防范风险。

a）技术措施：

1）搭建隔离的数据使用环境，专门构建了独立的数据分析平台，实现与外部网络的物理隔离；

2）实施严格的数据存储安全策略，包括全程加密存储、分散存储关键关联信息、建立定期备份机制；

3）构建精细的访问控制机制，支持基于角色的权限控制、多因素身份认证、全过程审计和与异常行为监测。

b）管理措施：

1）成立了专门的数据安全小组负责全流程监督，建立了完善的应急响应机制,对相关人员进行了严格的背景审查和安全培训，并实施关键岗位轮岗和双人复核机制；

2）与数据使用方签订了严格的协议，明确约定数据使用目的和范围，规定了违规使用的责任条款。

B.7 结果验证和效果评估

匿名化效果评估采用了多层次的评估方法，包括形式化验证和实证评估两大类。形式化验证主要通过数学模型验证数据是否满足预设安全标准，采用 K-匿名值结合流通场景与流通环境进行综合评估；实证评估则模拟潜在攻击者行为，通过链接攻击测试和同格攻击测试等方法评估重识别风险。此外，还进行了数据效用评估，验证匿名化后的数据可用性。

a）无法识别方面，结果显示该匿名化处理达到了预期效果。从可访问性角度看，所有直接标识符已经过删除和假名化，数据集中不含任何能直接识别个体的信息；从可关联性角度看，即使将此数据集与可公开获取的其他数据集进行关联，由于关键关联字段已经过泛化处理，成功关联的概率极低；从可指向性角度看，K-匿名性测试显示数据集中最小 K 值为 22，任何准标识符组合至少对应 22 条记录，无法精确指向单一个体；从流通范围角度看，本场景是组织外部两方的数据流通，属于受控公开共享，场景系数 S 可取 1/5；从环境保障角度看，已经搭建安全可信的数据

使用环境，且具备较为完善的管理保障能力，因此环境系数 E>1。综上所述，本场景的匿名化程度 A>4.4，超过预设目标（A>1），认定已达到匿名化要求；

b）不能复原方面，通过多种方法论证了数据的不可复原性。模拟攻击测试表明，由于数据在受控环境中使用，即使采用目前先进的技术，也无法从匿名化后的数据中复原出原始精确值。技术成本分析显示，尝试通过反向计算或穷举攻击方式复原原始数据的计算资源需求过高，在现有技术条件下不具备经济可行性。此外，具体复原不可能性分析显示：所有直接标识经过彻底删除，无技术手段可复原；时间信息泛化为时间段，使原始绝对时间无法准确推导；年龄泛化为年龄区间，使原始年龄值无法还原。综上所述，匿名化处理后的数据集在技术上已不存在合理的复原可能。