ICS 35.030
CCS L80

中华人民共和国国家标准

GB/T XXXXX—XXXX

数据安全技术
个人信息匿名化处理指南及评价方法

Data security technology —

Guideline and evaluation method for anonymization of personal information

（征求意见稿）

（本稿完成日期：2025 年 8 月 26 日）

在提交反馈意见时，请将您知道的相关专利与支持性文件一并附上

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

国家市场监督管理总局
国家标准化管理委员会　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。本文件由全国网络安全标准化技术委员会（SAC/TC 260）提出并归口。

本文件起草单位：清华大学、中国电子技术标准化研究院、蚂蚁科技集团股份有限公司、阿里云计算有限公司、深圳市腾讯计算机系统有限公司、北京快手科技有限公司、北京天融信网络安全技术有限公司、华控清交信息科技（北京）有限公司、郑州信大捷安信息技术股份有限公司、飞利浦（中国）投资有限公司、中国信息安全测评中心、上海计算机软件技术开发中心、中国信息通信研究院。

本文件主要起草人：金涛、王建民、周晨炜、张峰昌、白晓媛、李克鹏、孙勇、王昕、靳晨、刘为华、李世奇、王龑、高松、毛争艳、戚琳。

引　言

为贯彻《中华人民共和国个人信息保护法》关于“匿名化”的法律要求，规范个人信息匿名化处理活动，提升个人信息处理安全水平，特制定本文件。

本文件在现有国家标准基础上，重点参考了GB/T 37964—2019《信息安全技术 个人信息去标识化指南》、GB/T 42460—2023《信息安全技术 个人信息去标识化效果评估指南》等技术规范，并与GB/T 39335—2020《信息安全技术 个人信息安全影响评估指南》等标准相衔接，共同构成个人信息保护标准体系的重要组成部分。

本文件在现行标准基础上，聚焦“无法识别”与“不能复原”两大核心目标，提出不同场景下的参数建议，明确匿名化处理的完整技术路径与管理要求，配套提供对抗性测试方法与不能复原性核验机制，确保匿名化处理结果具备可审计性、可复现性与可验证性。

本文件适用于各类组织在开展个人信息匿名化处理、共享、发布等活动中的技术选型、过程控制、效果评价与合规审计，也可为监管部门、第三方评估机构提供技术参考与评价准则。

数据安全技术　个人信息匿名化处理指南及评价方法

1　范围

本文件明确了匿名化处理的目标，提供了个人信息匿名化处理的指南，并给出了匿名化评价方法。

本文件适用于个人信息匿名化处理工作，也适用于开展个人信息安全管理、监管和评估等工作。

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　信息安全技术　术语

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 42460—2023　信息安全技术　个人信息去标识化效果评估指南

3　术语和定义

GB/T 25069—2022、GB/T 35273—2020、GB/T 37964—2019、GB/T 42460—2023界定的以及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。

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。

3.3

匿名数据　anonymized data

匿名化处理后的数据。

3.4

数据属性标识度　data attribute identifiability degree

用于综合评价某一数据属性的内在唯一性与组合影响力，进而衡量该属性在重标识个人信息主体时风险贡献度的量化指标。

3.5

对抗性测试 adversarial testing

模拟目的明确的攻击者通过真实的数据收集和真实的外部数据资源，对已去标识化的数据集进行重标识攻击的一种测试。

注：攻击者具有一般人的合理技术能力，但不是具备特殊安全知识技能的专家。

4 概述

4.1 目标要求

匿名化建立在去标识化之上，是去标识化的特殊情形，目标是实现极低重标识风险与不可复原性。匿名化结果同时满足下列两项要求：

a) 无法识别：在数据接收方限定场景与设定环境风险下，结果数据不可识别特定自然人；
b) 不能复原：在合理可得的技术与资源下，结果数据不能被恢复为原始个人信息。

4.2 匿名化流程概述

匿名化处理的流程见图 1，实施过程包括：

a) 首先基于数据集特点和匿名化处理后数据接收方所限定的使用场景进行准备工作；
b) 进行个人信息去标识化处理并进行效果评估，直到处理后的数据集从数据接收方的角度评价达到 GB/T 42460—2023 中所述的个人信息标识度 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)；
c) 进行对抗性测试和不能复原性核验；
d) 对于未通过对抗性测试和不能复原性核验的数据集，需要重新进行去标识化处理及后续步骤；
e) 通过对抗性测试和不能复原性核验后，出具阶段性评价报告，作为内部质量把关说明；
f) 对匿名化处理的全过程进行有效的管理。

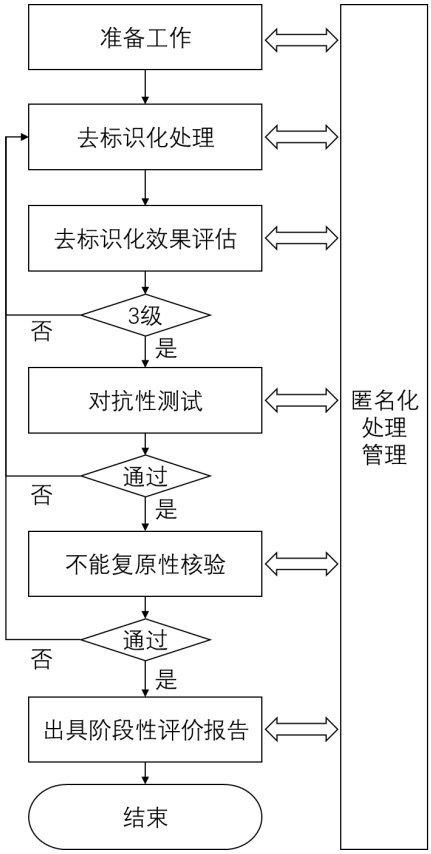

图 1　匿名化流程

4.3　匿名化评价概述

针对匿名化后的数据集以及匿名化处理过程包括管理体系，做正式判定并出具评价报告/证书，由独立角色（内部合规或第三方）完成，以确保匿名化结论成立，满足匿名化目标要求。从以下方面进行评价：

a)　无法识别的评价；
b)　不能复原的评价；
c)　对抗性测试评价；
d)　综合评价。

5　匿名化流程

5.1　准备工作

对于个人信息的匿名化处理，首先基于数据集特点和匿名化处理后数据接收方所限定的使用场景进行准备工作，包括：

a)　按GB/T 39335—2020开展目的、必要性与潜在安全风险分析；
b)　识别数据集特征（人群特征；类型：微数据/聚合/事务/时序等；属性类别：直接标识符/准标识符/敏感属性；统计特征：时间范围、样本规模等）；
c)　明确使用授权与使用场景（用途、授权范围与使用限制、隐私与安全要求等）；

d) 确定可接受风险水平与门槛参数（K-匿名、L-多样性、T-接近性、差分隐私参数 ε / δ 等）；

根据不同的共享场景，建议采用下列参数进行去标识化处理：

表 1 建议参数

维度	完全公开共享	受控公开共享	领地公开共享	说明
K-匿名（K）	≥20	≥5	≥3	或证明等效/更强保障
L-多样性（L）（如适用）	≥5	≥3	≥2	与 T-接近性择一或并行
T-接近性（T）（如适用）	≤0.05	≤0.10	≤0.20	距离度量可用 EMD 或等效证明
差分隐私 ε（如适用）	≤1	≤2	≤8	应同时给出 δ 并管理预算

差分隐私（如适用）优先用于聚合发布或经生成模型的合成数据；若用于微数据直接发布，应在报告中给出算法、 ε / δ 、组合规则与预算台账，并说明适用边界。

e) 制定处理方案与回炉（重新进行处理）条件。

5.2 去标识化处理

参照GB/T 37964—2019，选取隐私模型（K-匿名/L-多样性/T-接近性/差分隐私等）及技术机制（删除、泛化、抑制、分桶、扰动、微聚合、合成等），以GB/T 42460—2023标识度3级为最低目标。

根据不同的使用场景和数据提供方明确的隐私和安全要求，选择并部署相应的环境安全保障措施，包括但不限于访问控制、权限管理、安全审计等基本安全技术措施，以及可信执行环境、数据沙箱、安全多方计算等安全计算环境技术，以降低环境因素引发的重标识风险、复原风险。

5.3 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，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。

5.4 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，且尽可能地表现得像一个真正的入侵者。对攻击动机、攻击者来源、攻击者能力的模拟如下（具体攻击者模型见附录 A）。

1) 模拟攻击动机：动机一般分为目标明确和目标随机，包括但不限于对他人的恶意、经济利益与商业竞争，挖掘披露他人隐私，炫耀自身能力，政治或其他目的等。模拟攻击者模拟真实攻击者的动机进行模拟攻击，但不实施真正的个人伤害；

2) 模拟攻击者来源：包括但不限于朋友、同事、组织内部人员、公众、竞争对手、数据经纪商、广告商、记者、疏远的伴侣、商业间谍、研究人员、跟踪者等；

3) 模拟攻击者能力：覆盖弱/强对手模型，前提是使用合法可得的数据与合理可得的技术与资源；不得假设违法犯罪手段。

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（例如 0.6）的属性，重点分析其是否可作为用于攻击测试的关键变量，关键变量数据属性标识度量化计算方法见附录 B。

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：
   1) 开放的数据资源包括但不限于互联网、社交媒体、公共开放数据、图书馆等；
   2) 私有数据资源是匿名数据接收方的数据使用者可以合法获得的私有可访问数据资源；
   3) 着重考虑确切有记录的信息、既定的事实、常识性的背景知识等，不考虑特定的个人知识（但可以考虑个人职业身份相关知识）。
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，攻击方法包括但不限于：
   1) 记录匹配（精确/模糊）与跨库链接；
   2) 推理攻击/敏感属性推断；
   3) 成员推断；
   4) 去噪攻击（针对添加噪声或合成数据方案）。
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。

5.5 不能复原性核验

5.5.1 目的原则

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。遵循原则如下：

a) 核验应遵循最小必要、独立复核、证据留痕、可审计的原则；
b) 核验结论应基于客观证据与可重复的方法，并形成可追溯的记录；
c) 凡存在可逆映射关系或可获取之关键辅助材料导致可复原风险的，不得判定为通过；
d) 职责独立性：
   1) 组织应实行“双人复核”机制：处理责任人与核验责任人相互独立；
   2) 核验责任人不得参与具体匿名化处理实施与关键辅助材料的日常保管；
   3) 必要时可设第三方审计或内部合规部门进行抽查。

5.5.2 核验对象与范围

核验范围应覆盖生产、备份与灾备环境及可访问副本。核验对象应至少包括：

a) 原始数据结构及必要样本、处理环境、处理脚本、工具及处理日志；
b) 处理方法与参数（删除、泛化、抑制、分桶、扰动、微聚合、合成、差分隐私、哈希等）；
c) 关键辅助材料：盐、密钥、令牌映射表、随机种子、噪声主参数等；
d) 存储与计算环境、备份/快照、CI/CD（持续集成/持续部署）与代码仓库相关配置。

5.5.3 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。

a) 核验准备：
   1) 应形成《处理与材料清单》，列明数据资产、处理步骤、处理脚本与关键辅助材料的存放位置、管控人、生命周期等；
   2) 应形成《数据流拓扑图》，标识原始区、处理区、结果区及其隔离边界与访问路径等；
   3) 上述清单与拓扑图不得遗漏备份、快照、日志。

b) 技术不可逆性核验，对所用方法应逐项给出不可逆性说明与参数下限，包括但不限于：
   1) 哈希：应采用含随机盐的抗枚举方案，盐位数宜≥128 bit；对小域字段（如手机号、邮编、生日）宜叠加泛化/抑制；
   2) 令牌化/加密/格式保留加密：属可逆方法，不得作为终态匿名化手段；如出于流程需要暂存，后续应执行不可逆处理；
   3) 泛化/抑制/分桶/微聚合：应确保等价类规模与信息丢失程度满足设定阈值；小簇或过细分桶不得残留可逆推断路径；
   4) 扰动/加噪/合成数据：噪声或生成过程应与个体独立，随机种子不得复用；应给出噪声分布与尺度的合理性说明；
   5) 差分隐私（如适用）：应限定于聚合发布场景并给出 ε / δ 参数与预算台账，若用于微数据直接发布，应提供算法说明、 ε / δ 与组合规则、预算台账及适用边界证明。
c) 关键辅助材料不可得性核验：
   1) 盐、密钥、令牌映射表、随机种子、噪声主参数等应与结果数据在物理或逻辑上隔离；保留期限最小化，不得与匿名数据一起发布给数据接收方；
   2) 对于需要保存关键辅助材料的情况，关键辅助材料宜纳入KMS/HSM（密钥管理系统/硬件安全模块）或等强度管控，采用双人双控审批与全量审计；
   3) 如无合法合规的保留必要，关键辅助材料应按制度销毁并保留零化/销毁证据；如需保留，应明确期限、用途、访问人、审核流程与补偿控制，且由可信赖方妥善管理；
   4) 应对代码仓库、CI/CD、镜像与配置进行静态扫描，不得存在硬编码的 key/seed/salt；
   5) 备份与快照中不得保留可用于复原的关键辅助材料；历史版本如包含应删除或隔离并出具工单凭证。
d) 环境不可越权核验：
   1) 原始区与匿名结果区应网络/账户/权限隔离，采用最小权限与只读沙箱策略；
   2) 数据出域、参数变更与高风险操作应实行双人复核并留痕；
   3) 日志、中间表与缓存应定期清理；不得输出原始敏感字段或关键辅助材料；
   4) 应配置异常访问、批量导出、策略变更等监控与告警并定期核查。

5.5.4 判定结果

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。

5.5.5 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：

a) 处理方法或关键参数变更；
b) 新引入或撤销关键辅助材料；
c) 接收方环境风险变化或数据用途变化；

d) 发生疑似重标识或泄露事件；
e) 距上次核验超过12个月。

5.5.6 报告与记录

应形成《不能复原性核验报告》，至少包含：
a) 核验范围与对象；
b) 方法与参数、不可逆性说明；
c) 关键辅助材料隔离/销毁证据与审计日志；
d) 环境隔离与访问控制证据；
e) 例外与风险接受（如有）；
f) 结论与整改计划。

5.6 出具阶段性评价报告

5.6.1 职责与独立性

组织应实行双人复核机制：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。

5.6.2 输入与输出

输入与输出应至少包括：
a) 输入应至少包括：
   1) 数据与场景说明、使用授权与环境风险设定记录；
   2) 去标识化处理方案与参数、效果评估数据（K-匿名/L-多样性/T-接近性、等价类分布、差分隐私参数（如适用）及统计结果）；
   3) 对抗性测试方案、数据来源合规性、样本与结果；
   4) 不能复原性核验证据（技术不可逆说明、关键辅助材料隔离/销毁证据、环境隔离与访问审计）。
b) 输出应形成书面报告及证据包，并给出明确结论（通过/不通过/整改后复评）。

5.6.3 评价流程

评价流程应至少包括：
a) 资料完备性检查；
b) 去标识化效果结果审查；
c) 对抗性测试结果审查；
d) 不能复原性核验证据审查；
e) 形成结论并批准。

5.6.4 判定规则

判定规则如下：
a) 当下列条件同时满足时，结论应判定为“通过”：
   1) 去标识化效果满足既定门槛（如K-匿名/L-多样性/T-接近性、差分隐私参数，重标识风险）；
   2) 对抗性测试结果风险低于对应阈值；

3) 不能复原性核验证据充分，确认技术不可逆、关键辅助材料不可得、环境不可越权；
4) 资料完备、可审计、可复现。

b) 任何证据缺失、统计口径不当或关键辅助材料可得的情形，不得判为通过。

5.6.5 报告内容

报告内容要求如下：

a) 报告应包含以下内容：
1) 基本信息：项目名称、批次标识、时间、责任人/复核人、适用场景与环境风险设定；
2) 方法与参数：去标识化技术/模型与关键参数；对抗性测试设计（攻击者模型、开放/封闭世界、外部数据合规性）；不能复原性核验要点（技术不可逆、关键辅助材料不可得、环境不可越权）；
3) 指标与结果：K-匿名/L-多样性/T-接近性、差分隐私 ε / δ （如适用）、等价类分布与长尾处置，并给出统计结果；重标识风险；关键辅助材料（盐/密钥/随机种子/映射/噪声主参等）隔离或销毁证据；
4) 判定与处置：通过/不通过（回炉）；限制条件与整改要求；下一步安排；
5) 证据目录：参数与脚本（可复现实验，脱敏后）、执行日志与随机种子管理、KMS/HSM证据、审批与审计留痕。

b) 报告中的统计结果应同步给出方法、样本量、置信水平及关键假设。

5.6.6 批准与发布

批准与发布阶段要求如下：

a) 报告发布前应完成双人复核与电子/手写签署留痕；
b) 报告应进行编号与版本管理；发布范围与密级应与数据使用授权一致；
c) 对外提供报告摘要或指标时，不得披露可能降低保护强度的敏感参数（如具体盐值、密钥材料等）；
d) 报告与证据包应保存不少于三年或遵从更高法定期限；
e) 出现处理方法/参数变更、环境风险变化、数据分布漂移、发生安全事件，或自上次通过起满12个月的，应启动再评估并更新报告版本。

5.7 匿名化处理管理

匿名化处理管理包括以下内容。

a) 策略制定与实施：
1) 制定组织级匿名化原则、策略、流程与承诺目标；
2) 明确不同数据类型与场景的技术要求与参数上/下限；
3) 参照 GB/T 37964—2019 5.2 确定去标识化目标；
4) 如适用，建立差分隐私 ε / δ 预算台账、组合规则与熔断阈值，记录复用与分摊策略。

b) 组织与人员：
1) 明确负责部门/岗位与职责权限（参照GB/T 37964—2019 6.1）；
2) 定期培训与能力考核（参照GB/T 37964—2019 6.2）；
3) 建立外部共享/发布前的双人复核机制。

c) 不能复原性证据链：
1) 不可逆技术证明：哈希、扰动、泛化、微聚合、合成等技术与参数；
2) 关键辅助信息管理：盐、密钥、随机种子、噪声主参数等与原始数据物理或逻辑隔离；

3) 销毁或受控保留：如需保留，限定期限与访问控制；期满销毁并留存日志；
4) 复现实验与校验：提供可复现实验脚本的指纹/哈希与运行参数摘要（剔除秘密参数），并以审计日志证明一致性。

d) 风险管理与持续改进：
1) 建立匿名化风险台账与监控指标；
2) 监测数据环境与技术演进变化，必要时再评估与再测试；
3) 定期开展内部/外部审计，缺陷整改与追踪闭环；
4) 记录并保存关键环节日志与版本；
5) 再评估触发统一管理：处理方法/参数/场景或环境风险变化、数据分布漂移、事件、或自上次通过起满12个月，纳入年度复评计划。

e) 事件应急处置：
1) 制定并演练个人信息安全事件应急预案；
2) 发生事件时即时止损、报告与通知、原因分析与改进。

f) 合规与沟通：
1) 定期评估法律法规遵从情况；
2) 统一管理策略、制度、流程、评估与报告文档；
3) 与相关方沟通推广最佳实践。

g) 匿名数据的使用控制：
1) 审计访问与处理是否符合约定；
2) 当使用者、环境、目的变更时，及时通知发布方并评估影响。

6 匿名化评价方法

6.1 目标原则

匿名化评价作为正式验收/背书通常单列在匿名化处理过程（示例见附录C）之后，由独立角色（内部合规/第三方）完成并签署，匿名化评价报告可引用匿名化处理中的阶段性评价记录与证据包。

a) 以匿名化处理结果及管理证据为输入，包括数据与场景说明、环境风险设定、方法与参数、阶段性评价报告与证据包（脱敏后），形成通过/不通过结论并出具报告；
b) 宜由独立于匿名化处理团队的内部合规部门或第三方执行；第三方应出具无利益冲突与保密承诺；
c) 评价活动应遵循独立性、可复现、证据充分、最小必要披露与可审计原则；
d) 评价范围应明确数据边界、使用场景、环境风险设定与参数门槛；
e) 在不使用关键辅助材料（盐/密钥/随机种子/映射/噪声主参等）明文的前提下，复跑去标识化效果评估与对抗性测试脚本，并对应不能复原性核验开展证据审计与必要的抽样验证，核验一致性；
f) 评价结论应以书面报告与合规声明/证书形式发布，包含：管理摘要、项目与场景、限制条件、处理与评估、对抗性测试、不能复原性核验、综合结论、有效期、再评估触发与撤销条件、签署页、证据目录等；对外摘要不得披露降低保护强度的敏感参数（如盐值、密钥材料、随机种子等）；报告与证据包保存不少于三年；
g) 评价结论的有效期宜为12个月；当处理方法/参数/使用场景或环境风险设定发生变化、数据分布显著漂移、发生安全事件，或有效期届满时，应启动再评估并更新报告版本。

6.2 无法识别的评价

结合对抗性测试中发现的、在特定环境设定下可实际利用的风险进行综合判断，评价数据集在数据接收方场景下是否可以识别特定自然人，满足以下所有条件则通过无法识别的评价：

a) 环境风险设定合理：考虑访问主体（人数、角色、培训）、访问机制（网络/物理隔离、多因子）、存储与计算环境（专网/沙箱）、数据使用目的与期限、审计与追责能力等；给出设定理由、控制措施与残余风险说明；环境变化应触发再评估；
b) 结果数据不包含直接标识符；
c) 假名不可链接回直接标识符；
d) 准标识符满足K-匿名门槛要求；
e) 对于需要防止敏感属性推断的场景，敏感属性满足L-多样性或T-接近性门槛要求；
f) 采用差分隐私时，满足 ε / δ 参数与理论要求，并出具差分隐私预算台账。

6.3 不能复原的评价

结合对抗性测试中实施的、旨在复原原始信息或敏感属性的攻击测试结果进行验证，能够证明在考虑以下所有条件基础上，处理后的信息不能被恢复为原始个人信息，则通过不能复原的评价：

a) 覆盖统计推断、机器学习、外部匹配等攻击类型与弱/强对手模型；
b) 使用前沿方法（记录匹配、差分查询、机器学习、去噪攻击等）；
c) 测试场景贴近真实威胁，参考已知案例；
d) 多次运行与不同随机种子验证一致性；
e) 关键辅助信息不可得（隔离或销毁）且具备审计证据；如发现可逆路径或关键辅助材料可得，评价结论不得判定为通过；
f) 核验实施与证据要求见5.5。

6.4 对抗性测试评价

评价对抗性测试过程是否足够充分、有效，从以下几个方面进行衡量：

a) 测试设计的完备性：
   1) 攻击者模型覆盖的全面性：是否依据附录A，考虑了多种类型的攻击（如检察官攻击、记者攻击、营销者攻击），其动机、能力和可获取的背景知识是否与数据共享场景相匹配；
   2) 攻击方法与技术的代表性：测试所采用的攻击方法（如记录链接、推理攻击、成员推断、去噪攻击等）是否覆盖了当前技术条件下合理且前沿的攻击手段；
   3) 外部数据源的合理性：测试中所使用的外部辅助数据源是否模拟了真实攻击者可能“合法”获得的数据范围（如公开数据、商业数据库、社交媒体信息等），其广度、规模和相关性是否足以构成有效威胁。
b) 测试执行的严谨性：
   1) 测试的深度与广度：测试是否针对数据集中被评估为高标识度的属性（见附录B）以及潜在的脆弱点进行了重点攻击；测试的样本量或攻击尝试次数是否具有统计意义，能否可靠地评估重标识风险；
   2) 过程的可重复与可审计性：测试过程是否有详细记录（包括攻击脚本、使用的外部数据、测试步骤、随机种子等），使得测试在相同条件下可被复现和验证。
c) 结果分析的可靠性：
   1) 风险量化与解释的合理性：对测试结果（如成功重标识的记录数、推断出的敏感信息量）是否进行了准确的统计量化，并合理解释其残余风险水平是否低于预设的可接受风险阈值；

2) 迭代改进的证明：如果初轮测试未通过，是否基于测试发现的漏洞改进了匿名化方案，并进行了新一轮的测试，直至通过。

6.5 综合评价

满足以下所有条件则通过匿名化评价，评价结论应附参数表、统计结果、对抗性测试报告与完整证据清单。

a) 通过条件：
   1) 通过6.2无法识别的评价；
   2) 通过6.3不能复原的评价；
   3) 通过 6.4 对抗性测试的评价；
   4) 匿名化处理管理符合5.7内容；
   5) 任一关键指标不满足门槛阈值、证据不足或发现可逆路径/关键辅助材料可得的，不得判定为通过。
b) 审计清单与证据包括：
   1) 匿名化方案与回炉条件；
   2) 参数与脚本（可复现的去标识化与评估工具链）；
   3) 等价类与统计结果；
   4) 对抗性测试全套材料，包括数据来源合规性、方法、结果、失败样本与根因等；
   5) 关键辅助信息清单与隔离/销毁证据；
   6) 差分隐私预算台账（如适用）；
   7) 环境风险设定与控制措施；
   8) 内部/外部审计结论与整改跟踪；
   9) 事件响应与复盘记录。
c) 评价与测试报告包括：
   1) 项目概况，包括数据来源、规模、类型、时间范围、使用场景等；
   2) 处理方案与参数，包括技术组合、泛化层级、抑制比例、噪声分布、差分隐私参数与预算台账等；
   3) 效果评估，包括K/L/T结果、等价类分布、长尾处置等；
   4) 对抗性测试，包括攻击者模型、外部数据资源、方法、样本与统计判据、结果等；
   5) 不能复原性证据，包括不可逆说明、关键辅助信息隔离/销毁记录、访问控制与架构图、复现实验脚本/哈希等；
   6) 结论与建议，包括是否达标、限制条件、再评估触发点等；
   7) 附录，包括日志、版本、审计记录等。

附 录 A
（资料性）
攻击者模型

定义攻击者模型，需考虑多种因素，主要有攻击者本身的因素和影响攻击行为的因素，分别包括攻击动机、攻击者的来源、背景资源、攻击能力（知识能力和技术能力），以及攻击的目标、预期攻击效果、攻击影响范围等。

攻击动机包括恶意报复、经济利益/竞争、隐私挖掘、炫技、研究目的等。攻击者来源包括组织内部人员、公众、竞争对手、数据经纪商、广告商、记者、研究人员等。能力边界包括理解数据含义与变量类别，具备通用信息检索与数据处理能力，不包含专用黑客技能或违法手段。合法外部数据资源包括开放政府数据、统计年鉴、社交媒体公开信息、新闻/公告、图书馆、数据使用者依法可得的私有数据等。

依据ISO/IEC 27559:2022，典型攻击者模型可分为检察官攻击、记者攻击和营销者攻击等。

a) 检察官攻击：攻击者效仿检察官或调查人员在法律或司法程序中，寻找特定个体信息的行为。此种攻击者具备一定的背景知识（攻击者知道目标个体在数据集中），并且有动机去识别数据集中的特定个体。

b) 记者攻击：攻击者仿效记者在追求新闻真相、公共监督或报道独家故事时的行为。这种攻击者运用调查技能和资源，分析去标识化的数据集，以揭示和关联特定个体的敏感信息或个体身份（攻击者不知道或无法知道目标个体是否在数据集中）。

c) 营销者攻击：攻击者效仿营销者的行为，他们利用私有的或公开的身份数据库与去标识化数据集进行关联，以扩展对个体的多维度画像。这种攻击者的主要动机是商业利益，他们可能希望通过重新识别数据集中的个体（攻击者针对数据集中所包含的所有个体，而非某个特定的目标个体，试图识别数据集中尽可能多的个体），来定向推送广告、产品或服务，进而提高营销效率和营销策略的精确度。

用于判定攻击者模型类别的参考因素见表A.1。

表A.1 典型攻击者模型参考因素

参考因素	典型模型		
	检察官攻击	记者攻击	营销者攻击
攻击目标	特定人员	数据库内非特定人员	数据库内非特定人员
背景资源	知晓特定人员公开数据集	-	-
攻击能力	了解特定人员的身份属性信息	拥有私有的或者可公开访问的身份数据库,拥有较高重标识技术能力	拥有私有的或者可公开访问的身份数据库,拥有一定重标识技术能力
攻击动机	好奇特定人员的其他敏感属性信息	证明某人可以被重新标识,使得公开数据库的组织感到难堪或者名誉扫地	将额外信息与去标识化数据集关联,对身份数据库中人员的画像进行更多维度的扩展
攻击效果	关联并获得指定人员的个人信息	无指定人员,需证明重标识结果的正确性	无指定人员,无需证明重标识结果的正确性,仅需保证较高概率的关联性
其他因素	来自内部,仅对指定人员产生影响	来自外部,对个人信息主体影响有限	可能来自内部或外部,可能对大量个人信息主体产生影响
潜在攻击者	朋友、同事、组织内部人员等	公众、研究人员、竞争对手等	数据经纪商、广告商、黑灰产等

附 录 B
（资料性）
数据属性标识度计算

B.1 总体原则

对于目标数据集中的除直接标识符外的每个属性列可以量化计算其数据属性标识度。数据属性标识度的计算需要考虑每个属性列的独特性和影响力。数据属性标识度分值是每个属性列的唯一性分值和影响力分值的总和。

B.2 属性列唯一性分值计算

具有大量唯一值的属性列可以被认为具有很高的标识性。属性列的唯一性分值是指计算出的唯一记录的比率。如果唯一性分值为0，则认为该信息用于无法识别特定个体，不必视为准标识符。另一方面，具有非零唯一性分值的属性列意味着至少具有一个不同的取值，这样的属性列可以认为具有一定标识性，因为不同的取值可能帮助识别特定个体。可以参照下面的公式计算第i个属性列的唯一性分值。

$$S_u(i) = \frac{N_d(i)}{N_t(i)}$$

其中，$S_u(i)$是属性列i的唯一性分值，$N_t(i)$是数据列i的记录总数，$N_d(i)$是数据列i的唯一性取值记录总数。

B.3 属性列影响力分值计算

影响力是根据等价类数量的变化来衡量的。如果在排除特定属性列时等价类数量与整个数据集的等价类数量相比大幅减少，则说明该特定属性列对数据的标识性的影响力高。可以通过以下公式计算第i个属性列的影响力分值。

$$S_e(i) = 1 - \frac{N_E(T - C_i)}{N_E(T)}$$

其中，$S_e(i)$是属性列 i 的影响力分值，N_E是等价类函数，$T = \bigcup_{i=1}^{n} C_i$，是所有属性列的集合，$C_i$是第 i 个属性列。

B.4 数据属性标识度分值计算

使用唯一性分值和影响力分值的平均值评价属性列的数据属性标识度分值。

$$S(i) = (S_u(i) + S_e(i))/2$$

其中，$S(i)$是属性列i的数据属性标识度分值

附 录 C
（资料性）
匿名化处理过程示例

C.1 概述

本附录给出了医疗器械商收集医院的医学数字影像与通信(DICOM®)数据场景下的匿名化处理过程示例。

C.2 准备工作

C.2.1 数据收集的目的

为了验证新一代CT设备的图像质量，收集当前部署在某医院的CT设备所产生的胸部CT DICOM影像100例。

C.2.2 个人信息处理者

某医院。

C.2.3 匿名化数据接受者

某医疗器械生产商的CT研发团队成员3人。

C.2.4 数据特征

完整的DICOM医学影像包括了图像的像素信息以及描述影像检查的相关信息。DICOM的像素信息存储在DICOM Tag(7FE0,0010)中。描述CT检查的信息按照DICOM标准的格式存储在DICOM文件中对应的其它的DICOM字段中，例如：患者ID(0010,0020),患者姓名(0010,0010),患者出生日期(0010,0030),患者性别(0010,0040)[1]。

该案例中所收集的CT影像数据是胸部扫描数据，因此不包含颈部以上扫描数据。另外，除了DICOM数据文件外，也不收集与患者对应的任何其它结构化或者非结构化的数据。

C.2.5 潜在的风险

DICOM文件中存在直接标识符风险与准标识符风险。因为本次收集的数据不涉及颈部以上的CT扫描数据，因此不存在脸部及头部（数据重建）特征识别的风险。

直接标识符包含了患者本身的直接标识符（患者ID、患者姓名）以及DICOM各类数据实例对象的唯一标识符(instance UID),例如：SOP Instance UID, Study Instance UID, Series Insance UID etc.这些UID可以唯一定位到患者基本信息包括标识符。

准标识符主要包含了DICOM文件中存在的人口统计信息，如提到的出生日期、性别等。

C.2.6 可接受的风险水平以及匿名化方案

在本案例中，采用K-匿名隐私模型，K值设置为5，同时可接受的重标识残余风险不应超过0.03.该风险门限值的设定参考了GB/T 42460—2023附录D中推荐的总体风险(0.05)以及ISO/IEC 27559:2022附录B.2 Table B.2中的Non-public/Medium possibility of attack, medium impact(0.075)。

本案例拟采用K匿名隐私模型对DICOM影像数据进行匿名化处理。

C.3 去标识化处理

C.3.1 识别标识符

按照DICOM PS3.15 2024b附录E Attribute Confidentiality Profiles[2]识别标识符。

C.3.2 处理标识符

按照DICOM PS3.15 2024b 附录 E Attribute Confidentiality Profiles 表格 Table E.1-1 中的列Basic Prof.定义的Actions执行对DICOM中直接和准标识符的去标识化处理。虽然DICOM的元数据中相

1) [1] https://www.dicomlibrary.com/dicom/dicom-tags/

2) [2] https://dicom.nema.org/medical/dicom/current/output/chtml/part15/chapter_e.html

关的直接标识符和准标识符已经参照标准进行了完全的处理，但影像的像素数据中潜在的性别信息（依赖有经验的阅片人的技能）。因此在本案例中的去标识化阶段，性别仍然被作为准标识符。

C.3.3 验证审批

去标识化/匿名化专家组成员审查了匿名化方案、匿名化执行程序以及配置参数、运行日志等，并对匿名数据处理结果进行了抽检人工验证，结果符合预期。

C.4 去标识化效果评估

该案例的DICOM去标识化后的数据准标识符仅包含性别，因此数据以性别（男性、女性）分为2组。参照K匿名法，等价类的数量为2，分别简称为男性等价类，女性等价类。根据中国第七次人口普查，男性占比51.24%，女性占比48.76%[3]。我们在本参考案例中假定女性患者人数为40（及100位患者中女性占比为40%）。依K匿名法，重标识概率为$\frac{1}{K}$，其中k为等价类的大小，因此在本案例中男性、女性等价类的重标识风险分别是：$\frac{1}{60}$, $\frac{1}{40}$。那么：总体的重标识概率$\leq \max\left(\frac{1}{60},\frac{1}{40}\right)=\frac{1}{40}=0.025$. 根据GB/T 42460—2023的评估流程，将去标识化的DICOM数据集评定为3级。

C.5 对抗性测试

C.5.1 攻击者模型

该案例中，选择记者攻击作为攻击者模型。原因是本案例中数据接收方（3名来自某医疗设备厂商CT研发团队的成员)无法确认所要攻击(非恶意的不经意攻击)的目标人选在所采集的100例病人的DICOM CT扫描数据集中，因此检查官攻击不适合本案例。另外，该3名成员来自医疗设备厂商的CT研发部门，接受劳动合同以公司合规要求协议，营销者攻击也不适合该案例。

C.5.2 识别关键变量

在本案例中，经过去标识化后的DICOM数据中其它的元素都是图像本身以及检查的相关信息。在去标识处理中，经过处理后，不含直接标识符，性别作为唯一的准标识符，并得到了验证和审批。在对抗性测试中，将进一步分析残余风险。有经验的CT胸片阅读者很可能有机会从CT图像中识别出该图像对应的患者的性别。因此，该案例中将患者性别作为唯一的关键变量。另外，我们假设一个极端情况，在本案例中，某有经验的数据接收者查看了所有100例病人的CT胸部扫描，断定其中仅有一位女性。

性别的唯一性分值计算如下

$$S_u(i)=\frac{N_d(i)}{N_t(i)}=\frac{1}{100}=0.01$$

性别的影响力分值计算如下

$$S_e(i)=1-\frac{N_E(T-C_i)}{N_E(T)}=1-\frac{1}{2}=0.5$$

性别的数据属性标识度分值计算如下

$$S(i)=S_u(i)+S_e(i)=0.01+0.5=0.51$$

在本案例中选择标识度高于0.5的属性作为攻击测试的关键变量，性别的标识度为0.51,因此被选择作为攻击测试的关键变量。

C.5.3 攻击测试准备

在该案例中，我们不考虑外部其它非常规方式获得的辅助数据集用于攻击测试。本案例中考虑的用于辅助攻击的数据源有如下2种。数据源A)为3名研究者的熟人（个人关系网）；数据源B)为另一项研究所用的DICOM数据集。数据源A和B的进一步说明如下：

---

3) [3] https://www.stats.gov.cn/zt_18555/zdtjgz/zgrkpc/dqcrkpc/ggl/202302/t20230215_1904000.html

数据源A：3名研究者的熟人（个人关系网）。假定每个人的熟人为150人，总计为450人（假设男女占比各为50%）。数据源A包括的数据项有：姓名、性别、年龄、职业、居住城市。

数据源B：另一项研究所用的DICOM数据集。假定该3名研究者同时也在进行另一项研究，研究中使用了来自和本次案例所使用的同一家医院的DICOM数据。该数据集包含了200例患者（男女各100）的经过了去标识化后的DICOM。去标识化后的DICOM数据中包含患者年龄、患者性别、患者体重。假设其重标识风险为4.5%。

C.5.4 实施攻击测试

在本案例中，选择的攻击者模型为记者攻击。具体而言，就是测试确认该案例中性别为女性的唯一的那例DICOM数据能够关联到数据源A中的某条记录(对应到真实世界的一个可以唯一标识的人)的概率。测试分以下几种情况：

a) 使用性别关键变量将该案例的数据集和数据源A进行关联攻击测试(无其它辅助信息)。关联后，该案例中的性别为女的唯一的DICOM数据和数据源A中的225条记录（450中的女性，按照50%计算）产生了关联。如果没有其它辅助信息，在这种场景下，将该案例中的唯一女性DICOM数据关联到真实世界中的一个数据主体的概率不超过$\frac{1}{225} \cong 0.0044$.

b) 使用性别关键变量将该案例的数据集和数据源A进行关联攻击测试(有部分背景信息)。三名研究者中有人了解到自己的一名女性朋友近期（2个月内）去过该案例中所采集的DICOM数据的来源医院做过CT检查。此时，重标识的概率为单个患者的DICOM数据出现在本案例所收集的100例患者的概率。我们以单台CT单日检查量100（参考上海交通大学医学院附属瑞金医院北院 财政支出项目绩效评价报告 20184）计算2个月的总的CT检查量，$22 \times 100 \times 2 = 4400$，其中女性占比假设为50%，因此，该医院2个月内预计的女性CT检查总数为$\frac{4400}{2} = 2200$。因此，单个女性患者出现在该案例所收集的DICOM的数据集中的概率为$\frac{1}{2200} \cong 0.00045$

c) 使用性别关键变量将该案例的数据集和数据源B进行关联攻击测试。关联后，该案例中的性别为女的唯一的DICOM数据和数据源B中的100例女性患者的DICOM数据产生了关联。进行该关联的假设是本案例中的唯一女性患者的CT数据也出现在了同期进行的另一项研究所采集的数据(数据源B）中，假设概率为0.1。在这种场景下，重标识的概率计算需要分为2个步骤：首先与数据源B中某特定患者关联（以0.1的概率关联到数据源B)；其次，将特定的关联的数据源B中的记录与真实世界中的的数据主体进行关联。总的重标识风险为2个步骤的概率相乘的结果，即$0.1 \times 4.5\% \cong 0.0045$

C.5.5 评估攻击测试结果

在攻击测试中，涉及到了三种可能的攻击场景，其重标识概率分别是0.0044，0.00045和0.0045，均远小于所接受的风险水平0.03。另外，未发现任何其它可能的安全性问题。

C.6 匿名化处理管理

参照匿名化处理管理执行。同时也遵循案例中所涉及的医疗设备厂商的信息安全管理框架。

4) [4]https://www.shanghai.gov.cn/cmsres/ec/ecf81727244744c687f075b483f164f7/930b3a4aac9ae484c94b0b63f414d77b.pdf

参 考 文 献

[1] ISO/IEC 20889, Privacy enhancing data deidentification terminology and classification of techniques[S].

[2] El Emam K, Arbuckle L. Anonymizing health data: case studies and methods to get you started[M]. " O'Reilly Media, Inc.", 2013.

[3] Nelson, Gregory S. "Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification." SAS Global Forum Proceedings. 2015.

[4] Elliot, Mark, Mackey, Elaine and O'Hara, Kieron (2020) The anonymisation decision-making framework 2nd Edition: European practitioners' guide , Manchester. UKAN, 119pp.

[5] Information Commissioner’s Office (ICO). Anonymisation: managing data protection risk code of practice. 2012. Available from: http://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

[6] Jipmin Jung, Phillip Park, Jaedong Lee, Hyein Lee 0005, Geonkook Lee, Hyosoung Cha. A Determination Scheme for Quasi-Identifiers Using Uniqueness and Influence for De-Identification of Clinical Data. J. Medical Imaging Health Informatics, 10(2):295-303, 2020. [doi]

[7] ISO/IEC 27559:2022, Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework

[8] Wang, Tianhao, Jeremiah Blocki, Ninghui Li, and Somesh Jha. "Optimizing locally differentially private protocols." arXiv preprint arXiv:1705.04421 (2017).