ICS 点击此处添加 ICS 号
点击此处添加中国标准文献分类号

DB11

北　京　市　地　方　标　准

DB11/T ××××—××××

数据匿名化处理技术要求

Technical requirements for data anonymization processing

（征求意见稿）

（本轮合稿完成时间：2025 年 5 月 6 日）

××××－××－××发布　　××××－××－××实施

北京市市场监督管理局　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》规定起草。

本文件由北京市政务服务和数据管理局、北京市经济和信息化局提出。

本文件由北京市政务服务和数据管理局、北京市经济和信息化局归口管理并组织实施。

本文件起草单位：北京国际大数据交易所有限责任公司。

本文件主要起草人：

数据匿名化处理技术要求

1 范围

本文件规定了数据匿名化处理的技术要求，包括匿名化总体要求、匿名化处理流程、匿名化前准备、匿名化技术实施、匿名化后监测等阶段的要求。

本文件适用于以个人信息数据为主要对象的匿名化处理，不适用于非结构化数据。

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 35273 信息安全技术 个人信息安全规范

GB/T 37964 信息安全技术 个人信息去标识化指南

GB/T 39335 信息安全技术 个人信息安全影响评估指南

GB/T 42460 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

下列术语和定义适用于本文件。

3.1

数据 data

任何以电子或者其他方式对信息的记录。

[来源：GB/T 43697—2024，3.1]

3.2

个人信息 personal information

以电子或者其他方式记录的能够单独或者与其他信息结合识别特定自然人身份或者反映特定自然人活动情况的各种信息。个人信息经匿名化处理后所得的信息不属于个人信息。

[来源：GB/T 35273—2020，3.1]

3.3

去标识化 de-identification

个人信息等数据经过技术处理，使其在不借助额外信息的情况下，无法识别或者关联特定自然人的过程。

[来源：GB/T 35273—2020，3.15，有修改]

3.4

匿名化 anonymization

个人信息等数据经过技术处理，无法识别特定自然人且不能被复原的过程。

注1：个人信息经匿名化处理后所得的信息不属于个人信息。

注2：本文件中技术处理包括直接处理和受控开发。

[来源：GB/T 35273—2020，3.14，有修改]

3.5

标识符 identifier

数据记录中的一个或多个属性，在一个规定的语境中，能够用来唯一标识与其关联的事物的字符序列。

注：标识符分为直接标识符和准标识符。

[来源：GB/T 37964—2019，3.6，有修改]

3.6

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或数据相关描述对象的标识符。

[来源：GB/T 37964—2019，3.7，有修改]

3.7

准标识符 quasi-identifier

数据中的属性，结合其他属性可唯一识别特定自然人或数据相关描述对象的标识符。

[来源：GB/T 37964—2019，3.8，有修改]

3.8

数据集 data set

数据记录汇聚的数据形式。

注：数据集通常由数据记录（行）和数据属性（列）组成。

[来源：GB/T 5295—2017，2.1.46]

3.9

数据记录 data record

与某一描述对象、数据主体或事务相关联的一组信息。

3.10

数据属性 date attribute

数据记录描述的对象或实体的特征，表现为可以在数据集的记录中找到的信息类型，也称数据域、数据列或变量。

[来源：GB/T 18391.1—2009，3.1，有修改]

3.11

敏感属性 sensitive attribute

数据集（记录）中需要保护的属性，该属性值的泄露、修改、破坏或丢失会对数据主体产生损害。

[来源：GB/T 37964—2019，3.10，有修改]

3.12

等价类 equivalence class

数据集中所有准标识符属性值相同的记录行的集合。

[来源：GB/T 42460—2023，3.13，有修改]

3.13

匿名数据 anonymised data

经过匿名化处理后形成，在特定场景下，无法识别特定自然人且不能复原的数据。

3.14

可控安全环境 controlled security environment

在保障数据处理安全和风险可控的前提下，支持数据在不同系统、不同主体间进行提供、共享、计算等流通利用活动的软硬件设施集合。

注：在可控安全环境中进行数据处理，不增加数据所面临的安全风险。

3.16

重标识 re-identification

去标识数据被重新关联到原始数据记录描述的对象或实体的过程。

[来源：GB/T 37964—2019，3.9，有修改]

3.17

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—2019，3.12]

3.18

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1：例如通过协议禁止信息接收方发起对数据集中个体的重标识攻击，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—2019，3.13]

3.19

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—2019，3.14]

4 总体要求

4.1 合法合规

应遵守国家对数据安全和个人信息保护的有关要求，不得损害国家、社会和第三方组织及个人的合法正当权益，充分保障数据主体享受的法定权利。

4.2 安全优先

应将安全性作为数据匿名化处理的重要目标，采取充分的管理和技术措施，确保数据匿名化处理过程安全可控，并积极管控匿名数据后续流通使用行为，降低数据安全事故发生概率。

4.3 平衡效用

应根据业务实际和安全保护要求，面向场景化应用需求，选择恰当的匿名化处理方案，在满足合规目标和安全要求的前提下，提升匿名化后数据的可用性，促进数据安全性和可用性的有效平衡。

4.4 风险管理

应将风险管理贯穿数据匿名化处理全生命周期，事前组织数据处理影响评估，事中结合再识别风险判断匿名化效果，事后动态管理剩余风险，并定期开展风险评估，及时发现并处置可能暴露的各类风险。

4.5 过程可控

应采取充分的管理和技术措施确保匿名化处理过程可监控、可追溯、可解释，明确各方参与主体权责，规范各环节处理权限和流程，有效监控并记录处理过程，增强匿名化处理的透明度和可信度。

5 匿名化处理流程

数据匿名化处理流程应涵盖前、中、后三个阶段，确保数据安全风险动态可控，具体如图1所示：

a) 匿名化前准备，包括分析匿名化需求、明确匿名化主体、确定匿名化目标、划定匿名化对象、评估匿名化影响；
b) 匿名化技术实施，包括数据直接处理、数据受控开发；
c) 匿名化后监测，包括匿名化效果评估、匿名数据使用控制、定期安全风险评估。

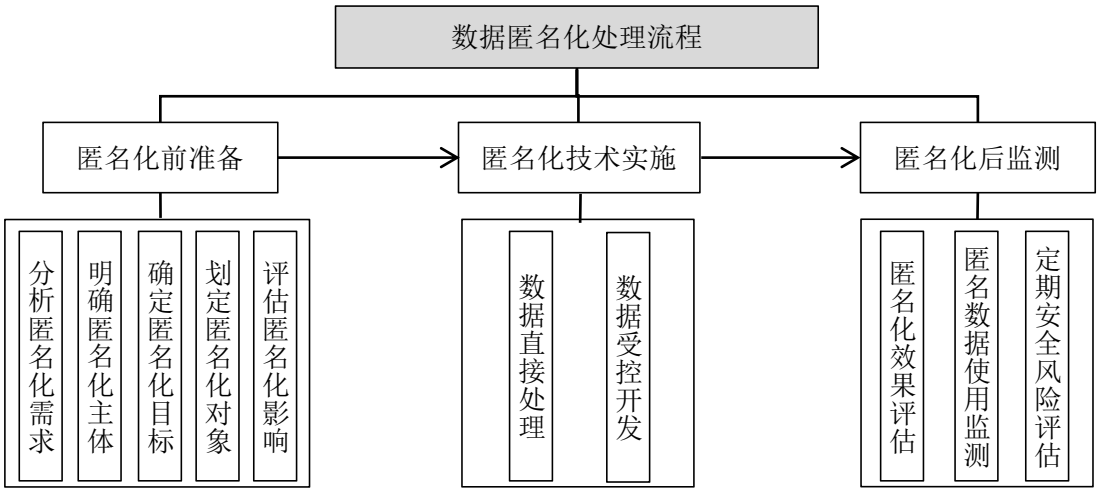

图 1 数据匿名化处理流程

6 匿名化前准备

6.1 分析匿名化需求

根据特定应用场景，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，通常需要考虑以下因素：

a) 匿名化处理行为是否正当合理；

b) 匿名化处理的执行成本是否可接受；
c) 匿名化处理后的数据是否可以支撑业务需求；
d) 匿名化处理后的数据安全风险是否可控制；
e) 匿名化处理是否存在其他可能的替代方案等。

6.2 明确匿名化主体

实施匿名化处理工作的主体包括组织方、执行方、监督方、使用方，多个角色可以由同一主体承担，相对复杂的匿名化任务，宜寻求专业机构或服务商支持。各方应签订承诺或协议，明确各自职责、权利和义务，规定处理标准和流程。

a) 组织方：负责组织匿名化处理实施，应对匿名化处理的行为和结果负责,确定匿名化的目标、范围和策略，并协调各方参与匿名化处理过程。组织方可以是数据提供方、数据处理者；
b) 执行方：负责具体执行匿名化处理操作，应具备专业的数据处理能力和丰富的实践经验，根据组织方委托的任务和目标，提供匿名化处理的具体解决方案和技术工具；
c) 监督方：负责对匿名化处理过程及结果进行监督、测试验证或审计，可以是独立的第三方机构、行业监管部门、专业组织，也可以是组织内部独立的监督审计、法务或风控部门；
d) 使用方：实际利用匿名数据进行研究、分析，应遵守数据安全和个人信息保护的有关要求，以及与组织方签订的相关承诺或协议，按照约定用途和范围使用数据。使用方可以是数据处理者。

6.3 确定匿名化目标

应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，相关考虑因素包括但不限于：

a) 原始数据的敏感程度；
b) 可采用匿名化技术的特性及成熟度；
c) 匿名数据的流通范围；
d) 可能的外部攻击风险；
e) 匿名数据后续的应用场景和业务需求等。

6.4 划定匿名化对象

明确需要进行匿名化处理的数据范围应遵循以下步骤。

a) 分析业务所需数据范围：应基于业务场景确定不同数据及各字段的匿名化需求，明确支撑使用方具体业务开展所需处理的数据类型和范围；
b) 判断可能关联识别的其他数据：应将组织方所控制的其他数据集或安全可控环境中，可能与上述数据具有关联识别可能的数据记录，一并纳入处理范围；
c) 分析待处理数据基本情况：应明确数据的性质、内容、格式、数量、相互关系、敏感程度和存储位置等。

6.5 评估匿名化影响

应结合所处理数据的敏感程度、拟采用技术方法的成熟度、匿名数据的使用方式和范围、组织机制的完备度、保护措施的充分性、可能的相关攻击等，事前进行影响评估。评估的内容包括但不限于：

a) 匿名化处理的目的、方式、范围等是否合法、正当、必要，以及对后续业务运行的作用和影响；
b) 拟采用的技术和方法的成熟度，以及是否有助于匿名化目标的实现；
c) 实施数据匿名化操作对组织内部业务流程、信息系统的执行需求及影响；
d) 针对匿名化处理过程中拟采取的监测、管控措施是否充分、有效并与风险程度相适应等；
e) 匿名数据的使用主体及其使用范围、安全防护能力等对限制重新识别难度的影响；
f) 匿名化处理可能对数据主体及其他利益相关方产生的影响；
g) 针对敏感个人信息的匿名化处理事前进行个人信息保护影响评估，评估应符合 GB/T 39335 的要求。

7 匿名化技术实施

7.1 概述

匿名化技术实施的路径包括数据直接处理和受控开发两种方案，前者要求处理后可直接输出匿名化的数据集，后者是对数据直接处理后仍残存一定风险或无法满足业务用途，但不涉及对外完全公开共享的，要求在可控安全环境开发使用，保证数据在可控安全环境中满足匿名化要求。附录A给出了部分行业领域特定业务场景的匿名化处理参考示例。

a) 数据直接处理：根据梳理应用场景中的数据范围，应用各类技术对范围中的数据直接进行匿名化处理，保证输出的数据集符合匿名化要求；
b) 数据受控开发：不涉及完全公开共享具体数据的，应在对数据进行基础处理后，将数据置于可控安全环境中进行开发利用，保证数据产品开发过程符合匿名化要求。

7.2 数据直接处理

7.2.1 数据标准化

将待处理数据进行必要的标准化，满足匿名化处理的数据格式规范要求。

a) 数据过滤：将非结构化数据和半结构化数据转换为结构化数据，对噪声数据进行平滑处理，对不符合业务规则的数据进行删除；
b) 错误标记：发现异常值或离群值，对残缺数据、失真数据、重复数据进行标记；
c) 修正处理：对已标记的错误数据采用针对性的方法进行处理，如对残缺数据进行内容补全，对逻辑冲突数据进行修正校验，对重复数据进行合并处理等；
d) 数据检验：对标准化后的数据进行检验，当检验结果未达到预期处理目标的，应再次进行数据标准化；
e) 数据加载：加载符合预期处理要求的数据，以进行后续的匿名化处理过程。

7.2.2 分类处理

根据数据的性质类别、可能识别数据主体的程度进行区分，分别确定技术措施，包括：

a) 含有直接标识符，或者结合公开数据集可识别出数据主体的；
b) 含有准标识符，或者结合其他数据集或准标识符可识别出数据主体的；
c) 不属于a）和b），但包含数据主体敏感属性的；
d) 应用场景中所必须使用的标识符或敏感属性，可单独处理。

7.2.3 技术措施

去标识化技术可以应用于匿名化，数据处理的相关技术措施可参考GB/T 37964的相关要求，包括统计、加密、抑制、假名化、泛化、随机化、数据合成等技术类别及K匿名、差分隐私等相关模型，不同技术方法也可组合使用，综合考量匿名数据流通环境、业务场景需求和数据属性特点，根据7.2.2以及各类技术产生的不同处理效果，组合形成不同的技术实施路径。

a) 根据后续匿名数据的流通环境，选取技术的策略包括但不限于：
    1) 如果匿名化处理后的数据应用于完全公开共享环境，至少采用统计技术对直接标识符及准标识符进行处理；
    2) 如果匿名化处理后的数据应用于受控公开共享环境，至少采用假名化技术对直接标识符进行处理，并采用加密等技术设置访问权限；
    3) 如果匿名化处理后的数据仅在领地范围长期存储处理，不涉及公开共享，至少采用加密等限制访问性的技术进行处理，并确保密钥安全。

b) 根据后续匿名数据使用的业务场景，选取技术的策略包括但不限于：
   1) 如果业务场景对数据完整性、关联度无保留要求，可采用抑制、假名化等技术对标识符进行处理；
   2) 如果业务场景对数据真实性、颗粒度无保留要求，可采用随机化、泛化等技术对标识符进行处理；
   3) 如果业务场景需要保留特定敏感属性进行分析，可采用统计、泛化、随机化等技术进行处理，在满足部分个人信息保护要求下输出；
   4) 如果业务场景属于无需保留原始数据真实性的开发测试场景，可采用数据合成技术，以人工方式创建的数据来代替原始数据。
c) 根据待处理数据类型的属性特点，选取技术的策略包括但不限于：
   1) 如果针对数值型敏感属性进行处理，可采用统计、抑制、泛化、随机化、合成等技术，如对收入状况采用泛化技术进行区间化处理；
   2) 如果针对日期型敏感属性进行处理，可采用抑制、泛化、随机化等技术，如对出生日期采用抑制技术通过统一的“*”进行屏蔽处理；
   3) 如果针对文本型敏感属性进行处理，可采用抑制、假名化、泛化、随机化等技术，如对姓名采用假名化技术以独立生成的“张三”、“李四”等代替真实姓名，对个人行为数据采用抑制技术进行特征标签化处理等；
   4) 如果待处理数据中存在极值、异常值、离群值等数据记录，可采用抑制、泛化等技术进行处理。

7.3 数据受控开发

7.3.1 适用情形

对经过处理后的数据集，仍残存较大安全风险或者无法满足业务可用性需求，仅在受控范围或领地范围内共享的，应构建可控安全环境，在环境中开发使用，保证数据产品开发过程及输出结果符合匿名化要求。可控安全环境能力要求见本文件7.3.2，相关技术参考附录B。

7.3.2 能力要求

可控安全环境能力要求包括以下内容：

a) 身份鉴别和访问控制：
   1) 应对各方身份进行鉴别，并利用数字签名等保证操作的抗抵赖性；
   2) 应具备对参与方进行权限管理和访问控制的能力；
   3) 应遵循最小化原则设置访问权限，定期对账号权限进行梳理，并及时清理过期和不合理权限。
b) 数据输入控制与安全存储：
   1) 应通过安全通道将数据传输至可控安全环境，采用加密、数字签名等技术对输入数据的机密性和完整性进行保护；
   2) 应对不同数据源的数据进行隔离存储，确保数据独立存储、权限独立管控；
   3) 应保证密钥、标识映射关系、日志记录等的存储安全。
c) 数据安全计算与高危行为拦截：
   1) 将数据计算严格限制在隔离的可控安全环境；
   2) 对受控开发过程中需要展示数据的环节进行动态脱敏，降低操作人员通过属性信息关联个人身份的风险；

3) 对受控开发过程中产生的衍生数据进行动态敏感字段识别，对产生的敏感数据进行处理；
4) 对操作行为进行实时监测与分析，对越权访问、遍历查询、关联或识别个人身份信息、批量导出数据等高危行为进行及时警告并阻断；
5) 宜保证受控开发任务的计算延迟、吞吐量、计算精度等满足数据处理业务的需求。

d) 数据输出控制：
1) 对数据处理结果应用实施数据利用鉴权，保证数据处理结果仅被指定的数据使用方获取，且应用方式和目的符合相关方协议约定要求；
2) 应保证数据提供方的数据不被安全环境外任何一方获取；
3) 除输出的数据处理结果外，应保证处理过程不泄露任何信息；
4) 应支持对数据处理算法逻辑的审查，确保其自身可解释性、安全性、可用性等要求，并将算法逻辑交由监督方确认、监督；
5) 数据处理完成后，应保证可控安全环境缓存的过程数据被安全删除。

e) 过程审计和数据溯源：
1) 受控开发过程中，应通过日志记录数据处理全流程并进行存证，包括主体、客体、时间、处理行为等，对数据链路进行刻画，且确保存证信息不可篡改；
2) 宜具备对存证和日志的安全审计以及相关处理行为的追溯能力，如采用水印技术等实现处理行为的可追溯性，并采用校验技术或密码技术保护溯源数据的完整性。

8 匿名化后监测

8.1 匿名化效果评估

8.1.1 评估维度

匿名化处理效果应从“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，在不借助额外信息情况下，是否满足使用方无法识别原始数据所描述的特定自然人的要求。相关评估流程可参考GB/T 42460；
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，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。

8.1.2 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 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) K-匿名值计算法：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。基于 K 匿名的效果评估方法可参考附录 C；
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) 风险排除法：结合 GB/T 39335 的要求，分析是否存在可预见的各类风险，判断对数据主体重识别或者对数据主体权益造成影响；
e) 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；

f) 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。

8.1.3 评估结论

应基于特定业务场景和数据处理环境，根据8.1.2中评估方法，对匿名化技术实施效果在8.1.1中各个维度进行评估。

a) 经评估，数据处理和受控使用后仍存在不可控的安全风险时，应在调整技术方案并重新处理后，再次进行评估；
b) 当业务需求或场景环境发生改变，需动态进行需求分析和风险识别，必要时重新进行处理；
c) 经多次方案调整和技术处理，其效果仍评估无法达到要求的，数据使用方应按GB/T 35273相关要求继续履行数据安全保护义务。

8.2 匿名数据使用控制

8.2.1 技术控制措施

应持续注意匿名数据的后续处理行为产生的风险，结合匿名化效果，配套不同的技术手段控制匿名数据的访问主体、内容、权限、载体，减轻可能的非法处理和滥用风险。

a) 访问主体控制：通过技术控制可访问匿名数据的主体范围，对在线查询或使用匿名数据的主体进行身份认证和鉴权，对离线使用的情况，通过密码保护或加密实现；
b) 访问内容控制：控制披露的匿名数据范围，仅提供匿名数据集的子集，且该子集可来自随机提取或经扰动处理；
c) 访问权限控制：对在线访问的数据仅开放匿名数据处理结果的查询功能，不开放匿名数据的保存、转发、打印等功能，并支持对数据访问权、使用权进行撤销；
d) 访问载体控制：匿名化程度较低时，可以要求访问主体到指定地点或使用指定设备处理匿名数据，限制将匿名数据带入其他环境，管理可能产生的相关设备链接。

8.2.2 管理控制措施

应采取充分的管理手段控制匿名数据后续处理行为，确保后续处理行为的目的、方式和范围符合规范。

a) 目的限制：审查使用目的的合法性，将匿名数据的使用限制在特定项目和用途中，要求接收方只能用于约定目的；
b) 最小必要：要求使用方应遵循最小化原则，在最小必要范围内对匿名数据进行处理；
c) 人员管理：各方承诺禁止反向识别，要求匿名数据使用者签订保密协议，定期进行合规培训和安全检查；
d) 密钥管理：妥善保管匿名化处理过程相关记录，安全保存有关密钥和映像表；
e) 转移限制：限制未经许可的匿名数据再转移，避免因扩散范围过广导致的多源串联交叉比对分析导致风险增加；
f) 共享记录：利用数据目录记录已共享数据集，防止不同数据集通过组合暴露敏感信息；
g) 及时销毁：定期清理已达成处理目的不再使用的匿名数据集，采取措施有效销毁任何意外重新识别的数据；
h) 应急处置：建立匿名数据处理安全事件应急处置机制，发现数据接收方、使用方等违反制度、

技术要求，或因其他原因导致数据泄露的，立即通知相关方停止处理行为，采取有效补救措施（如更改口令、回收权限、断开网络连接等）控制或消除面临的安全风险；

i） 违约责任：约定任何主体违反法律法规和相关协议约定需承担的严重后果。

8.3 定期安全风险评估

8.3.1 风险监测评估

为及时发现和响应潜在的风险，组织方应定期对匿名数据进行风险评估，并形成风险评估报告。评估的内容包括但不限于：

a） 匿名数据的处理主体、目的、方式和范围以及存储方式是否在设定框架内进行；

b） 针对匿名数据发布前遗留的剩余风险所采取的措施是否充分；

c） 匿名数据发布后是否产生了新的风险和隐患；

d） 数据匿名化效果不充分产生了哪些实际损害；

e） 法律环境变化带来的新合规要求是否对匿名化效果产生影响；

f） 技术环境变化产生的重新识别攻击手段带来的相关威胁；

g） 业务模式、应用环境变化带来的数据可用性需求对数据安全的影响等。

8.3.2 动态调整优化

组织方宜结合风险评估结论，追踪合规情况及与数据匿名化相关的最佳做法，动态更新和优化匿名化处理的技术和管理策略，确保数据匿名化处理持续符合相关要求。相关调整包括但不限于：

a） 使用更加严格、更符合实际的匿名化技术方法；

b） 调整匿名技术的参数以增强风险防范能力；

c） 实施更强有力的组织和管理措施；

d） 丰富匿名数据风险监测手段以及时发现新的攻击及相关隐患。

附 录 A
（资料性）
部分场景数据匿名化处理示例

A.1 金融行业合格投资人认证多方数据统计分析场景

A.1.1 处理场景

A.1.1.1 场景说明

为防止金融产品销售给不具备相应风险承担能力的个人金融投资者，监管部门要求金融机构销售金融产品时，必须严格对客户进行合格投资人认证（Qualified Investors），以此防范金融风险。获取客户在金融机构的资产情况越多，认证评估的准确性和真实性也就可能越高，降低业务风险的同时，也可为客户提供更适合的金融产品服务。因此销售金融产品的机构，期望查询其他机构的客户资产情况，并通过统计分析获得客户是否合格的评价。

A.1.1.2 需求分析

金融机构对外提供数据需要重新取得数据主体同意，但实际情况是：一是原本合规审查都是人工完成，无论对机构还是对数据主体，都费时费力；二是被查询金融机构的数据是基于存量业务累积，难以仅针对该场景单独向用户申请信息的查询使用。如果金融机构向第三方提供经过匿名化处理后的统计分析结论，则可以在保护数据主体权益前提下，降低机构和数据主体的成本，提高金融数据流通的效率。

A.1.2 处理对象

A.1.2.1 数据范围

本场景以涉及个人信息包括银行、信托机构和保险机构等需要处理的个人金融资产存量信息为代表，对个人投资者进行不暴露资产信息提供方机构名称、该个人在具体机构的金融资产存量数额的查询认证。在机构中存储的数据包括，客户的标识、标签（客户类型、所在地区等）、资产信息（资产额、资产估值等）。

A.1.2.2 数据性质

根据 JR/T 0171-2020 等标准，三要素（姓名、身份证号、手机号）、客户号、个人资产 C2 类别敏感数据，而三要素等更是直接标识符，客户号是准标识符。银行侧库表数据见表 A.1 所示。

表 A.1 银行侧库表数据

序号	字段名称	分类识别	标识符识别
1	客户号	C2 类别个人金融信息	准标识符
2	所属分行	C1 类别个人金融信息	非标识符
3	三要（姓名、身份证号、手机号）	C2 类别个人金融信息	直接标识符
4	资产余额	C2 类别个人金融信息	非标识符，敏感信息

A.1.2.3 处理目标

本场景中，对客户资产的统计分析结果越准确，既便于客户准确选择适当的金融产品，也降低销售金融产品的机构相关风险。基于最小化原则，仅使用该业务确需使用的数据，尽量降低数据可能造成的安全风险，再基于受控环境进行计算，并确保输出结果达到匿名化效果。

A.1.3 处理方式

A.1.3.1 技术方案

本场景选择数据直接处理的基础上进行受控开发的方案。根据第 7 章要求，对于客户号等业务不需要的数据进行删除处理，库表中剩余确需使用的数据，根据识别、敏感程度采用不同措施进行数据直接处理，再基于全密文环境进行受控开发，仅输出合格客户的判断结果。

A.1.3.2 分类处理

以表 A.1 中银行侧库表数据为例，数据直接处理方式见表 A.2。

表 A.2 银行侧库表数据直接处理

序号	字段名称	选择技术	数据处理方法
1	客户号	屏蔽	不需要使用，删除
2	所属分行	屏蔽	不需要使用，删除
3	三要（姓名、身份证号、手机号）	假名化/加密	需要作为核验索引，拼接后先进行密码杂凑处理，再使用公钥加密，以备隐匿信息检索
4	资产余额	加密	需要求和，使用公钥加密，以备半同态计算

直接处理后，以密文形式进行受控统计分析。

a) 通过密码协议隐匿信息检索（Private Information Retrieval）获取各金融机构的个人资产信息的密文。
b) 将各方密文输入到可信的第三方机构平台，使用半同态加密计算求和。
c) 查询方将其合格阈值信息（如-1000，代表 1000 万）加密后输入第三方机构平台，通过半同态加密将资产总额与合格阈值进行加法计算。
d) 输出结果到查询方，解密得到正数或负数结果（代表“是”或“否”）。

A.1.4 效果评估

A.1.4.1 处理效果

对受控开发后输出的结果数据进行处理效果评估。

a) 在无法识别方面，由于没有直接标识符在业务开展中使用，在全密文环境下获取和计算各金融机构的个人资产信息，因此输出的结果数据完全失去了与个人信息主体的关联和指向；但售卖金融产品的机构如果要将结果解密用于个人业务，需要根据用户同意开展业务；
b) 在不能复原方面，由于合格阈值保密，第三方无法获得用户的总资产；使用了 PIR 和半同态等密码算法，保障了对多方数据的统计分析计算的不可逆，输出结果数据无法反向推断用户在特定金融机构的敏感信息。

A.1.4.2 使用控制

在 A.1.3 的基础上，还需提供其他保障措施。

a) 基于合格投资者业务场景建立匿名化处理的策略和计划，确定安全管理负责人；

b) 确保半同态加密私钥单独存储，评估处理环境的风险，采取访问控制、安全传输和存储等措施；
c) 针对合格投资人认证业务，开展业务的金融机构间签订业务协议，明确补偿机制和退出机制，并可在隐私协议中告知客户；
d) 匿名化处理的过程保留数据存证和操作日志，如加密后信息的摘要信息等；
e) 业务达成后，保障密钥等安全销毁；
f) 如委托第三方机构协助处理资产信息，需要保证处理过程密文状态，以及监督第三方采用有效的安全保护措施；
g) 定期评估匿名化处理效果，并对处理效果进行持续监控。

A.2 互联网广告行业广告转化归因关联求交场景

A.2.1 处理场景

A.2.1.1 场景说明

本场景主要是广告行业中的广告归因场景，对广告平台的触点行为用户和广告主转化用户的数据进行关联求交。广告归因的过程，是确定哪个用户通过点击哪个广告完成了转化。通常后链路转化数据由广告主，或为广告主提供技术服务的平台所掌握。为此，广告归因需要双方数据合作完成。为最大限度的降低归因过程中的数据共享，降低双方的隐私顾虑，可以通过多方安全计算（Secure Multi-Party Computation）的方案，得出广告平台有触点行为用户和广告主转化用户的交集，进而实现归因。

A.2.1.2 需求分析

本场景中涉及的主体和利益相关方包括广告主和广告平台。广告主和广告平台在输出用户数据的过程中，对于广告平台侧的触点行为用户和广告主侧的转化用户的交集个数，虽然无法识别到个体用户，但是如果接收方获取到准确的广告归因用户数据，可以通过差分攻击的方式，套取对方的用户画像数据。因此，需要对广告平台侧的触点行为用户个数、广告主侧的转化用户个数、用户交集的广告归因数据进行保护。比如：广告主上传号码包 1，假设包含号码 1-100，按照学历维度查询，画像统计信息，得到：初中 20 人，高中 15 人，大学 30 人，硕士 14 人，博士 21 人。接着，广告主上传号码包 2，包含设备号 1-100 加 x，x 为广告主想要套取画像信息的设备号，广告主做同样的统计查询，得到：初中 20 人，高中 15 人，大学 30 人，硕士 15 人，博士 21 人。那么广告主可以知道在硕士中人数增加 1，那么就可以推断出增加设备号 x 是硕士，用户画像信息泄露。

为了进一步提升安全性，可以基于差分隐私技术，对用户画像数据进行保护。同时，对用户画像数据进行差分隐私保护时，添加噪声即可，计算成本较低，执行难度较低，也不会对归因效果，以及后续的业务用途产生影响，对数据可用性的影响程度较低。

A.2.2 处理对象

A.2.2.1 数据范围

需进行匿名化处理的数据包括：广告平台侧的触点行为用户个数、广告主侧的转化用户个数、用户交集的广告归因数据。

A.2.2.2 数据性质

广告平台侧的触点行为用户个数、广告主侧的转化用户个数、用户交集的广告归因数据均不涉及直接标识符和准标识符，属于敏感属性。因为触点行为用户个数、转化用户个数、交集用户个数属于商业机密。数据示例为整型数值，比如 10 个用户。

A.2.2.3 处理目标

对于广告平台侧的触点行为用户个数、广告主侧的转化用户个数，进行隐私求交处理。在隐私求交时，双方不需要知道对方的原始数据，只需要得到用户交集个数即可。对于用户交集的广告归因数据，进行差分隐私处理。差分隐私处理时，将“隐私损失”控制在一个特定水平内即可。

本场景不涉及对标识符的处理。

A.2.3 处理方式

A.2.3.1 技术方案

广告归因技术方案的流程如下：

a) 广告平台和广告主分别把触点人群 ID 和转化人群 ID 作为输入；

b) 广告平台和广告主执行可证明安全的 Circuit-PSI（Private Set Intersection，隐私求交）协议、得到 ID 是否匹配的交集信息，输出为布尔分片向量；

c) 广告平台和广告主根据布尔分片采用 MPC 的方式安全统计为真的个数，即交集个数，输出为交集个数的秘密分片；

d) 广告平台和广告主把秘密分片恢复成明文之前添加差分隐私，输出归因数据（即交集个数扰动后的数值）。

广告归因技术方案的架构图，如图 A.1 所示。

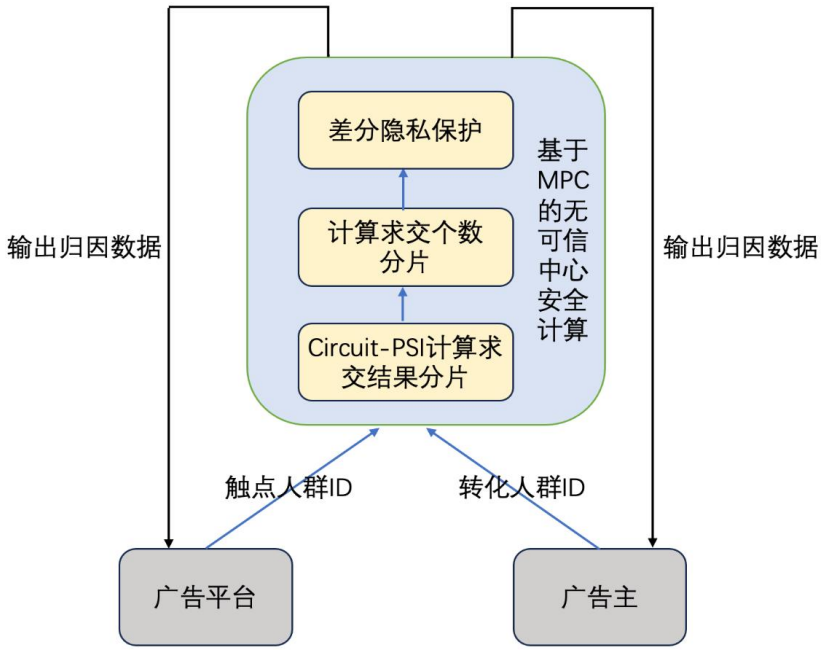

图 A.1 广告归因技术方案架构

本技术方案通过可证明安全的 Circuit-PSI 和根据布尔分片向量安全统计真值个数，一方面限制了数据的使用目的仅限于归因使用；另一方面，双方都无法接触对方的原始数据，最终只能得到广告归因的交集用户的统计数量，并非用户粒度的数据，最大限度地避免了用户个人信息的共享；最后，统计数量

经差分隐私保护，进一步提升了个人信息面对差分攻击时的安全性。

A.2.3.2 分类处理

针对广告平台侧的触点行为用户个数、广告主侧的转化用户个数，执行可证明安全的 Circuit-PSI（Private Set Intersection，隐私求交）协议、得到 ID 是否匹配的交集信息，输出为布尔分片向量；广告平台和广告主根据布尔分片采用 MPC 的方式安全统计为真的个数，即交集个数，输出为交集个数的秘密分片。

针对用户交集个数，广告平台和广告主把秘密分片恢复成明文之前添加差分隐私，输出归因数据（即交集个数扰动后的数值）。比如，用户交集个数为 10，添加差分隐私之后，输出为 9。

A.2.4 效果评估

A.2.4.1 处理效果

本示例采用拉普拉斯机制，将服从拉普拉斯分布的随机噪声添加到用户交集个数上。由于添加噪声对原始数据进行了微小的扰动，添加噪声后的数据会被控制在预设的隐私预算内，使得攻击者无法识别出准确的原始数据，从而实现“不可识别性”。另外，由于拉普拉斯分布的特性，噪声的大小与数据敏感度成正比，使得即使攻击者知道数据的统计方法和噪声分布，也难以从添加噪声后的数据中准确还原出原始数据，从而实现“不可还原性”。

对于差分隐私的效果，可以通过 Pr{数据泄漏风险}=Pr{计算结果反推原始数据风险} + Pr{计算过程泄漏数据风险} * Pr{场景下攻击可能性}来计算。在上述广告归因场景下，假设输入 A 和 B 的输入规模均为 n，计算结果为 k < n。由于整个计算是在可证明安全的 MPC 协议下执行，因此，Pr{计算过程泄漏数据风险}约等于ε，一个非常小的数，即在约定敌手攻击能力下，无论其发送何种攻击，数据泄漏的概率都非常小。加上差分隐私之后，反推原始数据的概率至少下降了ϵ/2 倍，满足匿名化要求。

A.2.4.2 使用控制

在受控环境层面，双方采用安全多方计算的环境，双方都无法接触对方的原始数据，最终只能得到广告归因的交集用户的统计数量，并非用户粒度的数据，最大限度地避免了用户个人信息的共享。

对于差分隐私中剩余风险的控制，可以通过调节隐私预算的方式来实现。隐私预算决定了差分隐私机制引入的噪声量。可以根据实际的应用场景中，数据的敏感度以及隐私保护需求，动态调整隐私预算。如果数据泄露风险较高，可以设置较小的隐私预算以提供更强的隐私保护。

A.3 医疗健康行业个性化降压用药联合建模场景

A.3.1 处理场景

A.3.1.1 场景说明

在临床治疗高血压场景中，通常是从已证明安全有效的降压药物中随机选择降压药物，但是不同降压药物存在巨大的个体差异，遗传因素、环境因素和健康状况与高血压的治疗效果也密切相关。因此可以采取个性化降压方法，为每个患者选择特定的降压药物类别，而非根据目前建议从临床试验中已证明安全有效的降压药物中随机选择降压药物。此场景中，需要用已有海量患者的相关数据进行用药模型训练，来支持临床医生的用药决策。

A.3.1.2 需求分析

为实现个性化降压方案，需要大量的高血压患者相关遗传信息、环境信息、健康信息以及用药信息对用药模型进行训练，但由于模型训练涉及数据量大、范围广泛、时间跨度长，因此，需要采用匿名化方案构建医疗用药模型以辅助医生用药决策，以充分保护个人信息主体权益。匿名化影响评估如下：

a) 该场景中需要用到大量患者的家族史、身体质量指数（BMI 指数）、所在地区、用药前后血压值等数据，这些数据的收集都是满足合法性基础的，并且数据相对敏感度低，比如用药前后血压值一般是动态的，对于标识个人的作用较小；
b) 该场景中的联合建模会采用集中枢纽方式，通过一系列管控措施，切断枢纽内数据和外部数据的关联，外部数据进不来，内部数据出不去，能够有效控制重识别风险、数据泄露和滥用风险；
c) 匿名化处理过程中拟采取的监测、管控措施充分、有效，能够有效防止重识别风险、数据泄露和滥用风险；
d) 匿名数据用于辅助医生用药决策，范围可控，能够有效防止重识别风险、数据泄露和滥用风险；
e) 匿名化处理不会对数据主体产生负面影响。

A.3.2 处理对象

A.3.2.1 数据范围

支撑个性化降压用药联合建模需进行处理的数据范围涉及高血压患者相关遗传信息、环境信息、健康信息以及用药信息，具体包括以下数据类型：

a) 患者社保ID：字符串型；
b) 是否有家族史：布尔型；
c) BMI指数：数值型；
d) 运动频率：数值型；
e) 是否饮酒：布尔型；
f) 职业：文本型；
g) 所在地区：文本型；
h) 所用药物：文本型；
i) 用药前血压值：字符串型；
j) 用药后血压值：字符串型。

A.3.2.2 数据性质

待处理的各类数据具有下列属性：

a) 直接标识符：患者社保ID；
b) 候选准标识符：所在地区、职业、运动频率；
c) 敏感个人信息：是否有家族史、BMI指数、所用药物。

A.3.2.3 处理目标

在满足安全条件和合规要求的前提下，为支撑联合建模顺利进行，各类数据的处理程度宜符合以下条件。

a) 是否有家族史：需要准确数据，但该属性不具备唯一性，对于重识别的效果有限；
b) BMI指数：需要保留整数位，或者给出区间值，以判断是否超重；
c) 运动频率：需要至少给出区间值，且区间值不宜过于宽泛；

d) 是否饮酒：需要准确数据，但该属性不具备唯一性，对于重识别的效果有限；
e) 职业：需要给出大致的范围，以判定是否压力过大等；
f) 所在地区：泛化到省级单位即可；
g) 所用药物：需要精确用药信息支撑建模；
h) 用药前血压值：需要至少给出区间值，且区间值不宜过于宽泛；
i) 用药后血压值：需要至少给出区间值，且区间值不宜过于宽泛。

A.3.3 处理方式

A.3.3.1 技术方案

数据匿名化处理技术方案主要包括模型训练阶段和决策阶段。

在模型训练阶段，将训练医疗用药模型的多维海量数据进行去标识化处理后加密上传至可信受控环境，先进行机器学习训练，得到医疗用药模型，并且在模型训练的前、中、后期均确保匿名化，做到“可算不可识”。在决策阶段，将高血压患者的遗传信息、环境信息、健康信息等进行去标识处理，上传至可信受控环境，输入到预先训练好的模型中，最终得到个性化降压用药方案以支持医生决策。

A.3.3.2 分类处理

数据先期处理方式如下。

a) 患者社保ID：用HMAC算法对ID进行假名化处理，再用受控环境的公钥对假名进行假名处理；
b) 是否有家族史：用受控环境的公钥进行加密处理；
c) BMI指数：先进行泛化处理，再用受控环境的公钥进行加密处理；
d) 运动频率：先进行泛化处理，再用受控环境的公钥进行加密处理；
e) 是否饮酒：用受控环境的公钥进行加密处理；
f) 职业：先映射替换为大的职业类别，再用受控环境的公钥进行加密处理；
g) 所在地区：先映射替换为大的区域，如东北、西北、华南等，再用受控环境的公钥进行加密处理；
h) 所用药物：用受控环境的公钥进行加密处理；
i) 用药前血压值：先进行泛化处理，再用受控环境的公钥进行加密处理；
j) 用药后血压值：先进行泛化处理，再用受控环境的公钥进行加密处理。

受控环境构建方式如下，具体如图 A.2 所示：

图 A.2 个性化降压用药联合建模受控环境构建方式

a) 数据加工方建立一个受控环境，对该受控环境设置严格的身份鉴别、访问控制、数据流入流出等方面的管控措施；
b) 参照GB/T 37964-2019，对用于建模的大量用户数据的标识符进行去标识化处理，确保去标识化后的数据在受控环境中，不结合额外信息的情况下无法被重标识；且受控环境密钥由可信第四方管理；
c) 通过安全通道将数据传入数据融合计算方的受控环境，并在其中完成计算；
d) 如果计算结果中仅包含统计信息，可以输出到受控环境外部；如果通过多个统计信息可能反推出个人身份信息，则需在输出前采用差分隐私、k匿名等进行处理；
e) 如果计算结果中包含个人信息，在征得相关个人同意后方可以输出到受控环境外部。

A.3.4 效果评估

A.3.4.1 处理效果

从匿名化处理后数据的可识别性和复原可能两方面分析处理效果：

a) 无法识别方面，在数据传输进可信受控环境之前，对所有数据进行加密处理，确保数据在传输过程中的机密性；数据在可信受控环境中进行建模训练，可信受控环境进行严格管控，切断与外部数据的关联，对于可访问可信受控环境的人员也进行严格管控，数据进行去标识化处理，确保了在可信受控环境中的数据无法识别特定自然人；
b) 不能复原方面，对于可逆性的去标识化技术“假名化”处理的数据“患者社保ID”，用 HMAC算法对ID进行假名化处理，且密钥由第四方持有，数据处理方无法获得密钥，故不能对数据进行还原；其他数据均采用了不可逆的泛化技术，且泛化后的BMI指数、运动频率、血压值等属于不具备唯一性的属性，即这些属性是动态变化的，不能对数据进行有效还原；职业和所在区域泛化后，由于在受控环境中使用，也不能结合其他信息进行推断还原。

A.3.4.2 使用控制

从访问控制、高危行为拦截、安全审计、数据溯源、应用安全等方面提出匿名数据后续使用行为的相关要求。

a) 访问控制要求包括：
   1) 基于数据分类分级配置访问控制策略；
   2) 访问控制粒度达到主体为用户级或应用级，客体为接口级、应用级、数据库字段级，确保未经权限审批的用户和应用无法访问数据；
   3) 权限设置遵循最小化原则，定期对账号权限进行梳理，并及时清理过期和不合理权限；
   4) 定期对账号进行梳理，及时清理过期账号、闲置账号、非法账号；
   5) 对数据处理活动中使用的工具，如数据脱敏工具等的数据访问和操作进行权限控制；
   6) 对主体、客体的访问控制策略和权限等进行统一管理。
b) 高危行为拦截要求包括：
   1) 对融合数据的访问和操作行为进行实时监测与分析，对识别出的风险，如越权访问、关联或识别个人身份信息、批量导出数据等进行及时告警，并能够阻断；
   2) 对接口遍历、数据爬取等行为进行实时监测与分析，对识别出的风险进行及时告警。
c) 安全审计要求包括：
   1) 实现对数据融合处理全流程的完整记录，包括主体、客体、时间、处理行为等，对数据链路进行刻画，实现可追溯、可审计；
   2) 数据加工方对匿名化处理过程进行记录并定期审计。

d) 数据溯源要求包括：
   1) 采用技术手段，如水印技术等实现敏感数据在数据展示时的可追溯性，并采用校验技术或密码技术保护溯源数据的完整性；
   2) 根据文件、应用及数据库表的操作日志和访问日志，构建用户、数据、代码解析等的关系刻画，获取字段与字段间的访问与调用关系（如继承、聚合、转换等），实现字段级的血缘刻画；
   3) 基于数据血缘管理能力，实现敏感数据溯源和操作行为溯源，降低数据安全风险。
e) 应用安全要求包括：
   1) 对数据处理结果应用实施数据利用鉴权，结合相关方协议约定，确保获得数据权利关系方许可同意，并且应用方式和目的满足协议约定要求方可应用；
   2) 对于包含群体挖掘结果的场景，对数据进行加噪后满足差分隐私或 K 匿名等保护要求下进行应用；
   3) 实现数据应用的流转管控措施，确保数据应用被准确记录和管理，及时阻断高危数据应用；
   4) 采用安全通道、通道加密、数据加密等措施保护数据传输安全，如：HTTPS、VPN 等，并采用国家认可的加密算法，如：SM2、AES 等。

A.4 医疗健康行业专病临床数据科学研究场景

A.4.1 处理场景

A.4.1.1 场景说明

针对医院神经外科专病临床数据库的颈动脉支架手术数据集和颈动脉剥脱手术数据集，通过匿名化处理，在满足患者隐私信息保护和医学研究伦理合规的前提下，将匿名化后的数据用于支持颈动脉相关术式手术技术的科学研究，助力医疗机构改进手术技术、优化治疗方案。

A.4.1.2 需求分析

颈动脉手术相关研究需要足够大的样本量才能保证研究结论的科学性，单纯依靠获取患者同意新增病例的方式难以在合理时间内积累足够数据。国家鼓励科研机构、高等学校、医疗机构、企业根据自身条件和相关研究开发活动需要开展人类遗传资源保藏工作，并为其他单位开展相关研究开发活动提供便利。开展涉及人的生命科学和医学研究时，使用匿名化的信息数据开展研究可免除伦理审查。匿名化处理后，数据不再属于个人信息，可以合法用于科学研究，同时保障患者隐私不受侵害，实现了法律合规和研究需求的平衡。同时匿名化处理会对数据精确度、完整性和关联性产生一定影响，但不会显著降低数据对目标研究的支持价值。

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结构化数据，不涉及复杂的非结构化或动态更新内容，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（如手术方式、治疗效果等）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特定研究团队内部，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。

A.4.2 处理对象

A.4.2.1 数据范围

本场景数据匿名化处理范围主要涵盖颈动脉手术患者的临床数据。一方面，这些数据直接支持颈动脉手术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) 患者基本信息：姓名、性别、年龄、身份证号、联系方式、住址、工作单位、婚姻状况、医保信息、住院号等。这类信息是典型的个人标识信息，直接关联到特定个体，必须进行严格匿名化处理；

b) 临床诊疗信息：入院诊断、出院诊断、病史记录（包括既往史、家族史）、体格检查记录、实验室检查结果（血液、生化、凝血等）、影像学检查结果（超声、CTA、MRA、血管造影等）、手术记录（手术时间、术式、操作者、手术过程等）、药物使用记录、护理记录。这类信息既包含研究所必需的临床特征，也包含可能间接识别患者的信息；

c) 随访及结局信息：出院后随访记录、并发症信息、预后和转归。这类信息对评估手术长期效果至关重要，同时也可能包含特定时间点和特殊情况记录，需要谨慎处理以避免识别；

d) 此外，虽然业务场景不直接涉及，但可能关联导致识别的其他数据包括：就诊时间、入院及出院时间、主诊医生、责任护士信息、特殊病例讨论记录、医院科室信息等。这些辅助信息与外部数据结合可能导致患者被识别，因此也被纳入匿名化处理范围。

A.4.2.2 数据性质

结合数据集的使用场景，对颈动脉手术临床数据集中的直接标识符、准标识符、敏感属性进行识别，具体表 A.3 所示：

表 A.3 颈动脉手术临床数据示例

原始数据示例	数据属性类型	数据字段
姓名：张三，男，65 岁，身份证号：110101195701XXXXX 住址：北京市 XXXXXXXX，联系电话：139XXXXXXXX 职业：退休教师，婚姻状况：已婚 入院日期：2023-03-15，出院日期：2023-03-28 主诊医师：张医生，手术医师：李医生 诊断：右侧颈内动脉重度狭窄 手术方式：右侧颈内动脉剥脱术 手术日期：2023-03-18 术后情况：恢复良好，无神经系统并发症	直接标识符	姓名
		身份证号
		医保号
		住院号
		详细住址
		电子邮箱
		病历号
	准标识符	性别
		出生日期/年龄
		入院/出院日期
		手术日期
		区/县级地理位置
		职业信息
		婚姻状况
		身高体重
		特殊社会身份（如军人、公务员等）
	敏感属性	颈动脉剥脱的具体诊断及病情严重程度
		合并症和并发症信息
		手术治疗详细记录
		预后情况
		特殊治疗（如实验性治疗）记录
		家族遗传病史
		精神状态评估

A.4.2.3 处理目标

为保障临床数据的最低风险且满足后续科学研究的最低使用要求，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，根据数据流通范围（限定研究机构内部使用）和数据敏感程度（医疗敏感数据）的综合评估，将风险控制目标设定如下：
   1) 识别概率阈值≤1/10000，基于数据交易流通环境中的安全等级要求，结合《个人信息安全规范》中对敏感数据的保护要求确定，确保任何个体在匿名化数据集中的识别概率低于1/10000；
   2) K 值≥5，基于数据集样本量（1.5 万余例）和准标识符组合种类（约 10 类）分析确定，确保任意组合的准标识符至少对应 5 个不同记录；
   3) L 值≥3，根据敏感属性（如诊断类型、手术方式等）的多样性分布统计结果，结合业内常见做法确定，确保每个等价类中敏感属性至少有 3 种不同值；
   4) 满足 T-接近度要求，确保每个等价类中敏感属性分布与整体分布足够相似，降低属性推断风险。
b) 为保证匿名化处理后的数据可满足后续科学研究的最低要求，将数据的可用性目标设定如下：
   1) 保留疾病特征与治疗结局的关联性；
   2) 保留时序性治疗过程的分析价值；
   3) 保留年龄、性别等与疾病相关的人口学特征；
   4) 确保数据可用于统计分析和模型训练。
c) 基于上述目标，对相关标识符及敏感属性的取舍要求设定如下。
   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：年龄按 5 岁区间分组（如 60-64 岁，65-69 岁）；入院/手术日期转换为相对时间（如首次就诊后第 N 天）；地理信息仅保留地级市信息，不显示详细地址；职业信息按大类分组（如教育工作者、医疗工作者等）。
   3) 敏感属性：保留疾病严重程度、治疗方式及效果等核心医学信息；特殊情况（如罕见并发症）通过泛化处理，避免特异性识别。

A.4.3 处理方式

A.4.3.1 技术方案

在专病临床数据科学研究场景下，单纯的直接处理可能会过度损失数据价值，特别是对于需要精确分析的医学研究，单纯的受控开发又需要较高的技术和资源投入，因此可选择数据直接处理结合受控开发环境的综合方案进行匿名化，确保基础安全性的同时，最大化保留数据研究价值，并可以根据不同情形的安全级别和研究需求，提供差异化的数据访问服务。

A.4.3.2 分类处理

先期数据直接处理包括去除所有直接标识符，对准标识符进行泛化处理，对特殊敏感属性进行抑制或概括，具体如下。

a) 直接标识符分别采用以下技术方法处理：
   1) 姓名、身份证号、联系方式：抑制处理，完全删除；
   2) 病历号：假名化处理，替换为研究编号（如 Bio20230001），保证患者具有唯一 ID；
   3) 其他需要保留但不可识别的标识符：使用加盐散列技术进行散列化处理。
b) 准标识符处理分别采用以下技术方法处理：

1） 年龄：泛化处理，由精确数值改为 5 岁区间（如 61-65 岁）；
2） 日期：泛化处理，转换为相对时间（如入院后第 X 天）或按月/季度粒度处理；或随机化处理，对入院日期增加-7 至+7 天的随机偏移，保持相对时间关系不变；
3） 地理位置：泛化处理，仅保留市级信息；
4） 职业：泛化处理，分类为大类（如教育类、医疗类、行政类等）。

c） 敏感属性处理分别采用以下技术方法处理：
1） 分级泛化：将具体疾病描述泛化为严重程度分级，将详细手术过程概括为手术类型和关键步骤；
2） 数值微扰：对实验室检查结果增加微小随机扰动（±5%），保持异常值标志和临床意义；
3） 局部抑制：对极为罕见的病例特征（发生率<0.1%）进行抑制处理，保留主要临床特征，抑制可能导致识别的独特特征。

d） 受控开发环境构建方式如下：
1） 搭建隔离的数据分析环境,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；实施严格的访问控制和行为审计，记录全过程审计日志并建立异常行为实时预警机制；
2） 设置差分隐私查询接口频率和精度要求，实施差分隐私机制为查询结果添加随机噪声，限制最小查询粒度以禁止单条记录查询，并设置查询频率限制，防止通过多次查询推断个体信息；
3） 输出结果在导出前需经专人审核，须经过风险评估后方可导出，确保不包含可识别信息。

原始数据与匿名化后数据对比如表 A.4 所示：

表 A.4 原始数据与匿名化后数据对比

原始数据	匿名化后数据
姓名：张三，男，65 岁，身份证号：110101195701XXXXX 住址：北京市 XXXXXXXX，联系电话：139XXXXXXXX 职业：退休教师，婚姻状况：已婚 入院日期：2023-03-15，出院日期：2023-03-28 主诊医师：张医生，手术医师：李医生 诊断：右侧颈内动脉重度狭窄（狭窄率 97%），合并高血压（160/95mmHg） 手术方式：右侧颈内动脉剥脱术 手术日期：2023-03-18 术后情况：恢复良好，无神经系统并发症	研究 ID：Bio20230089，性别：男，年龄段：65-69 岁 地区：北京市，职业类别：教育工作者，婚姻状况：已婚 入院时间：基准期+73 天，住院天数：13 天 诊断：颈内动脉狭窄（单侧），狭窄程度：重度；合并心血管疾病：高血压（二级） 手术类型：颈内动脉剥脱术 手术时间：入院后第 3 天 术后结局：恢复良好，无并发症

A.4.4 效果评估

A.4.4.1 处理效果

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，包括 K-匿名性、L-多样性和 T-接近度评估；实证评估则模拟潜在攻击者行为，通过背景知识推断测试、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，验证匿名化前后的统计特性保持度和关键医学分析结论的稳定性。

a） 在无法识别方面，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，所有直接标识符已完全删除，数据集中不含任何能直接识别个体的信息；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 7，超过预设目标（K≥5），任何准标识符组合至少对应 7 条记

录，无法精确指向单一个体；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；从可推断性角度看，背景知识推断测试表明，即使攻击者知道患者的性别、年龄段和所在地区等多项信息，识别成功率仍低于 0.01%，远低于风险阈值；从可区分性角度看，经过处理后的数据记录之间差异性减小，难以通过特征组合区分出特定个体；

b) 在不能复原方面，通过多种方法论证了原始数据的不可复原性。模拟攻击测试表明，即使采用目前先进的数据恢复技术，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：直接标识符经过彻底删除，无技术手段可复原；时间信息转换为相对值且添加了随机偏移，使原始绝对时间无法准确推导；地理位置信息泛化为市级单位后，每个地理单元对应大量人口，无法复原到具体地址；数值型检测结果添加了随机微扰，使原始精确值无法还原；特殊罕见特征通过局部抑制处理，防止了特征组合导致的唯一识别。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。

A.4.4.2 使用控制

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剩余风险。

a) 技术措施
   1) 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、建立定期备份机制和实现外部网络物理隔离；
   2) 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、全过程审计和异常行为监测；
   3) 设置严格的查询限制，禁止返回小样本查询结果并实施差分隐私保护。
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。

A.5 医疗健康行业电子处方询价隐匿查询场景

A.5.1 处理场景

A.5.1.1 场景说明

医院患者诊疗后形成药品电子处方，基于数据匿名化处理，在充分保护数据和隐私安全的前提下，患者可选择通过匿名化处理的方式，在确保自身信息、用药信息安全的前提下，将匿名化处理后的电子处方数据与药品提供方的供药清单数据进行线上交互比对，隐匿查询药品供应及价格信息，以确定前往哪个药品提供方处取药能够实惠、便捷地取到所有药品。

A.5.1.2 需求分析

电子处方询价场景涉及的主体和利益相关方主要包括医保电子处方流转平台、药品提供方和患者。该场景下重点需要对患者提供的个人电子处方数据进行匿名化处理。电子处方中的药品清单与患者所患疾病紧密关联，药品清单中的信息泄露可能威胁患者的隐私安全。匿名化处理后的信息无法识别特定主体且不能复原，采用匿名化处理技术可以平衡隐私保护与业务效能，既防止患者敏感信息泄露，又确保

药品供应方能基于匿名化数据协助完成询价，通过混淆算法、同态加密、零知识证明等技术，在保障计算过程中患者与药品提供方隐私安全的前提下，由药品提供方进行药品供应匹配，判断是否满足患者用药需求，实现药品的供需匹配、询价。

A.5.2 处理对象

A.5.2.1 数据范围

需进行匿名化处理的数据范围为患者电子处方上承载的各类信息，具体包括以下数据类型：

a) 患者姓名：文本型；
b) 证件号码：字符串型；
c) 疾病信息：文本型；
d) 药品信息（药品通用名、药品剂型、药品规格等）：文本型、数值型；
e) 药品使用剂量：数值型；
f) 药品需求数量：数值型。

A.5.2.2 数据性质

待处理的各类数据具有下列属性：

a) 直接标识符：患者姓名、证件号码；
b) 准标识符：药品信息、药品使用剂量、药品需求数量；
c) 敏感属性：疾病信息。

患者电子处方上承载的各类信息示例见表 A.5：

表 A.5 患者电子处方数据示例

患者姓名	证件号码	疾病信息	药品通用名	药品剂型	药品规格	药品使用剂量	药品需求数量（板/盒/瓶）
张 XX	11XXXXXXXXXXXXXX1	冠心病	硝酸甘油片	片剂	0.5mg×100 片/瓶	0.5mg/次，胸痛时舌下含服	1
			阿托伐他汀钙片	片剂	20mg×7 片/板	20mg/次，每日 1 次，晚间服	4
李 XX	11XXXXXXXXXXXXXX2	慢性阻塞性肺疾病	噻托溴铵粉雾剂	吸入粉雾剂	18μg×10 粒/盒	1 粒/次，每日 1 次，晨间吸入	3
			氨溴索口服溶液	口服溶液	100ml:0.3g/瓶	10ml/次，每日 3 次，餐后服用	2

A.5.2.3 处理目标

为实现电子处方数据的最低风险且满足后续用途的最低要求，对相关标识符及敏感属性的取舍要求设定如下。

a) 完整性方面：电子处方的药品供需匹配只需使用电子处方数据中的药品信息、药品使用剂量、药品需求数量等数据，因此完整性方面，电子处方数据在匿名化处理后保留药品信息、药品使用剂量、药品需求数量等字段即可，患者姓名、证件号码等身份信息和疾病信息无需保留；
b) 颗粒度方面：药品信息、药品使用剂量、药品需求数量等准标识符相结合仍然存在识别到特定主体的可能，因此需对数据进行进一步的处理，结合场景对电子处方数据的低颗粒度需求，通过进一步减少各属性的唯一值，降低数据颗粒度，从而增加识别特定数据主体的难度，保障电子处方数据安全。

A.5.3 处理方式

A.5.3.1 技术方案

电子处方询价场景的数据处理方式是在对数据进行直接处理的基础上进行受控开发。先期采用抑制、泛化等数据匿名化处理技术对电子处方数据进行处理，处理后的电子处方数据中仍然包含药品信息、药品使用剂量、药品需求数量等属性，属于准标识符，因此后续处理采用受控开发的数据处理方式。

A.5.3.2 分类处理

A.5.3.2.1 基于隐私计算技术，构建数据匿名化处理的受控开发环境，在数据匿名化处理的各个环节，针对不同的数据采用不同的处理技术。先期数据直接处理如下。

a) 电子处方数据直接标识符与敏感属性抑制：删除身份信息、疾病信息，保留药品信息、药品使用计量与药品需求数量，形成药品需求清单；
b) 药品需求清单泛化：通过药理和病理分析混淆的方式向真实药品需求清单中混入难以直接区分的混淆药品，构造混淆药品需求清单；
c) 基于 OT 的数据库查询：在数据库查询中采用 OT（不经意传输）技术，基于混淆药品需求清单进行药品的存在性查询、药品规格匹配、药品库存匹配、药品总价询价；
d) 同态加密计算：结合同态加密技术进行药品品类匹配计算、药品需求量匹配计算、药品总价计算。

A.5.3.2.2 可控安全环境构建方式如下，具体如图A.3所示：

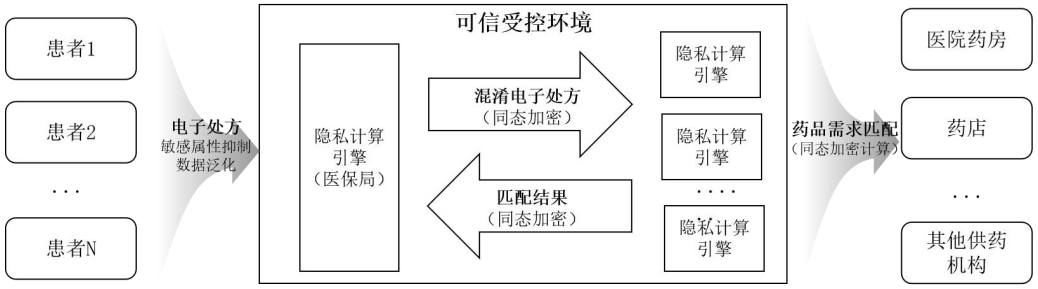

图 A.3 电子处方询价隐匿查询可控安全环境构建方式

a) 技术服务商构建可信受控环境，为各参与方在本地部署支持同态加密算法、混淆算法、不经意传输算法等算法的隐私计算引擎；
b) 将患者电子处方数据抑制、泛化处理后注入可信受控环境，基于隐私计算引擎的同态加密算法实现与供药机构的药品数据匹配计算；
c) 供药机构完成需求匹配后基于可信受控环境返回是否满足需求的匹配结果和药品总价，患者可查看满足需求的供药机构名称、地址及药品总价等结果信息。

A.5.4 效果评估

A.5.4.1 处理效果

本场景中，患者的电子处方为需匿名化的隐私数据，面向药品提供方输出的匿名化数据为去除了患者身份信息、疾病信息并对真实药品信息进行匿名化处理的混淆药品清单，在保障交互双方隐私不泄露的前提下实现供需精准对接。匿名化的结果数据消除了直接标识符，包含部分准标识符，重标识风险低于可接受风险值。

a) 在无法识别方面，电子处方数据经抑制处理后删除了患者身份信息这一直接标识符与疾病信息这一敏感属性，降低了数据的可指向性、可关联性，并对剩余属性的唯一值进行泛化处理，进一步降低了数据的可关联性和可推断性，有效提高了匿名化数据再识别的难度，且在可信受控

环境中对匿名化数据进行同态加密处理，确保匿名化数据在可信受控环境中无法再识别关联到具体患者；

b) 在无法复原方面，电子处方数据匿名化处理遵循“最小必要”原则，仅保留药品询价所需的核心属性，通过抑制技术删除敏感字段，泛化技术模糊化数据，同时在管理上严格访问控制，经模拟攻击测试，确保攻击者难以通过暴力破解或关联分析还原原始数据，关联到具体患者。

A.5.4.2 使用控制

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剩余风险。

a) 技术措施

1) 基于访问主体控制技术，对电子处方询价主体进行身份认证，确保使用电子处方进行询价的是患者本人，保障患者隐私信息安全；

2) 基于访问内容控制技术，确保电子处方询价主体仅获取自己的电子处方询价结果，保障药品提供方供药清单安全。

b) 管理措施

1) 电子处方询价场景对各参与方进行管理，审查使用合法性，医保局与技术服务商签署协议，保证数据基于隐私计算技术使用，只能用于电子处方询价场景，并承诺禁止进行反向识别，同时在协议中约定任何主体违反法律法规和相关协议约定所需承担的责任与后果；

2) 对不再使用的匿名数据及时进行销毁，同时建立匿名数据安全事件处置机制，对数据泄露进行及时处理，采取有效补救措施控制或消除负面影响。

附 录 B
（资料性）
可控安全环境技术参考

B.1 通用安全技术

通用安全技术可参考表B.1。

表B.1 通用安全技术

技术名称	应用场景	实现效果	提供匿名化保障
身份鉴别	通用	鉴别用户身份，防止非法用户登入系统	保障只有合法用户才能进入安全环境
访问控制	通用	对用户权限进行管控，防止非授权用户访问或操作系统资源	保障只有授权用户才能在安全环境内进行操作
安全传输	通用	确保数据传输过程中的机密性和完整性	保障数据传输中无信息修改或泄露
安全存储	通用	确保数据存储过程中的机密性和完整性	保障数据存储中无信息修改或泄露
安全监测	通用	监测和感知数据处理全流程中安全风险	为数据安全风险的控制和应对措施提供支撑
安全销毁	通用	确保数据处理后无信息残留	保证数据彻底清除且无法恢复

B.2 隐私保护技术

隐私保护技术可参考表B.2。

表B.2 隐私保护技术

技术名称	应用场景	实现效果	提供匿名化保障
多方安全计算	多方数据融合	确保数据加工、使用过程中的机密性和完整性	在安全模型有效的前提下，保障数据处理过程中全密文进行，无法被识别和恢复，且处理结果仅为指定参与方获得
同态加密	外包计算	确保数据加工、使用过程中的机密性	保障加密数据在第三方处理过程中全密文进行，无法被识别和恢复，且处理结果仅为密钥持有者获得
机密计算	多方数据融合通用	确保数据加工、使用过程中的机密性和完整性	在硬件隔离环境中，保障数据处理过程中无法被识别和恢复
分布式机器学习	联合建模	确保联合建模过程中，可泄露的信息最小	通过交互的中间数据满足统计上的泛化、抑制，达到个体无法识别的效果
差分隐私	联合统计、联合建模输出	确保联合统计、建模结果输出的隐私性	通过结果输出的泛化处理，保障个体无法识别，降低对个体的影响
零知识证明	验证共享数据真实性	确保验证过程的机密性，	保障验证过程无需暴露任何信息

		且仅授权方可验证数据	

B.3 流通控制技术

流通控制技术可参考表B.3。

表B.3 流通控制技术

技术名称	应用场景	实现效果	提供匿名化保障
智能合约	通用	确保数据按预定策略处理	保障数据按约定使用
区块链	通用	确保数据处理过程可信、可记录、可追溯	保障数据使用的时候追溯和审计
数据水印	通用	确保数据共享可追溯	保障数据使用的事后追溯
数据沙箱	数据开放	防止非授权用户处理和导出数据	在隔离环境中，保障数据按约定使用，且指定方获得结果
数据使用策略	数据共享	确保数据按预定策略共享和使用	利用可编辑策略，保障数据按约定使用

附 录 C
（资料性）
基于K匿名的效果评估方法

C.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C.1：

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

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，表示数据集中，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，如数据长期存储、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。

C.2 数据集的K匿名值

对于给定的离线库表结构的数据集，其K匿名的最小K值的计算步骤如下：

a) 先识别出直接标识符和准标识符；由于评估方与匿名化处理方通常不同，因此需要重新识别；

b) 针对匿名化处理后的准标识符的组合，找出数据集中所有的等价类；即数据集中与某个数据项具备相同的准标识符的数据项的集合；

c) 针对每个等价类，找出等价类大小的最小值，即K-匿名中的K值。

C.3 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C.2。

表C.2 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。

C.4 环境系数

对于环境保障能力，宜采用定性评估的方式。环境保障能力的中间值建议为1。当环境保障能力较低时，建议调高对于K匿名值的要求，以提升整体的匿名化程度。环境系数应对技术实施阶段的受控环境和匿名化后监测阶段的安全防护水平进行综合评估，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；重识别攻击的动机和能力，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。

参 考 文 献

[1] GB/T 18391—2009 信息技术 元数据注册系统(MDR)
[2] GB/T 35295—2017 信息技术 大数据 术语
[3] JR/T 0223—2021 金融数据安全 数据生命周期安全规范
[4] GB/T 43697—2024 数据安全技术 数据分类分级规则
[5] 中华人民共和国网络安全法（2016 年 11 月 7 日第十二届全国人民代表大会常务委员会第二十四次会议通过）
[6] 中华人民共和国个人信息保护法（2021 年 8 月 20 日第十三届全国人民代表大会常务委员会第三十次会议通过）
[7] 征信业务管理办法（2021 年 9 月 17 日中国人民银行 2021 年第 9 次行务会议审议通过）
[8] 汽车数据安全管理若干规定(试行)（2021 年 7 月 5 日国家互联网信息办公室 2021 年第 10 次室务会议审议通过）
[9] Personal Data Protection Commission Singapore, Guide to Basic Anonymisation. 31 March 2022.
[10] Article 29 Data Protection Working Party (European Commission), Opinion 05/2014 on Anonymisation Techniques. 10 April 2014.
[11] European Data Protection Board, Guidelines 01/2025 on Pseudonymisation. 16 January 2025.