

ICS 11.020

CCS C 07

T

团体标准

T/CEATEC XXX—2026

自身免疫病多组学数据库建设及应用 技术要求

Technical requirements for construction and application of multi-omics

databases for autoimmune diseases

（征求意见稿）

2026-X-XX 发布

2026-X-XX 实施

中国欧洲经济技术合作协会 发布

目 次

前言 VI

引言 VII

1 范围 1

2 规范性引用文件 1

3 术语和定义 2

4 总体原则 4

 4.1 依法合规 4

 4.2 临床与科研需求驱动 4

 4.3 统一标准 4

 4.4 最小必要 4

 4.5 分层管理 4

 4.6 安全内生 4

 4.7 全程可追溯 4

 4.8 原始数据保护 4

 4.9 可扩展与互操作 4

 4.10 质量可量化 5

 4.11 责任可落实 5

 4.12 持续改进 5

5 组织治理与职责 5

 5.1 治理体系 5

 5.2 角色与职责 5

 5.3 数据来源单位责任 5

 5.4 委托处理与共同处理 5

 5.5 制度与文件 5

 5.6 培训与授权 5

 5.7 利益冲突与监督 6

6 总体架构 6

 6.1 架构组成 6

 6.2 业务架构 6

 6.3 数据架构 6

 6.4 应用架构 6

 6.5 技术架构 6

 6.6 安全架构 6

 6.7 部署方式 6

 6.8 环境隔离 6

6.9 高可用与扩展 6

6.10 架构文档 7

7 数据来源与接入 7

7.1 数据来源登记 7

7.2 临床数据 7

7.3 样本数据 7

7.4 多组学原始数据 7

7.5 处理后和分析数据 7

7.6 外部数据 7

7.7 接入协议 8

7.8 安全传输 8

7.9 身份映射 8

7.10 接入验收 8

7.11 增量更新与撤回 8

8 数据标准与数据模型 8

8.1 数据分类与分层 8

8.2 患者主索引 8

8.3 样本标识与关系 8

8.4 临床数据模型 8

8.5 多组学数据模型 8

8.6 元数据模型 8

8.7 数据字典 9

8.8 术语和代码集 9

8.9 单位和精度 9

8.10 时间与事件模型 9

8.11 缺失和未知状态 9

8.12 完整性约束 9

8.13 模型变更管理 9

9 数据预处理与入库 9

9.1 总体流程 9

9.2 暂存区 9

9.3 文件与语法校验 9

9.4 数据清洗 9

9.5 重复和冲突 10

9.6 异常值 10

9.7 组学预处理结果 10

9.8 质量状态标识 10

9.9 入库审核 10

9.10 入库日志 10

9.11 回滚与重新处理 10

9.12 数据冻结与发布 10

10 存储与数据生命周期 10

10.1 分层存储 10

- 10.2 原始与衍生数据分离 10
- 10.3 容量管理 10
- 10.4 数据完整性 10
- 10.5 备份 11
- 10.6 恢复验证 11
- 10.7 灾难恢复 11
- 10.8 保存期限 11
- 10.9 归档 11
- 10.10 删除与销毁 11
- 10.11 服务退出和数据迁移 11
- 11 版本、血缘与可追溯管理 11
 - 11.1 数据集版本 11
 - 11.2 参考资源版本 11
 - 11.3 软件与流程版本 11
 - 11.4 模型版本 11
 - 11.5 数据血缘 12
 - 11.6 人工操作留痕 12
 - 11.7 审计追踪 12
 - 11.8 更正、撤回和影响分析 12
 - 11.9 可复现性验证 12
- 12 数据质量管理 12
 - 12.1 质量管理体系 12
 - 12.2 质量维度 12
 - 12.3 质量目标和阈值 12
 - 12.4 质量规则 12
 - 12.5 来源质量 12
 - 12.6 样本与组学质量关联 12
 - 12.7 运行质量监测 13
 - 12.8 问题分级 13
 - 12.9 不合格数据 13
 - 12.10 质量报告 13
 - 12.11 质量抽查和复核 13
 - 12.12 多中心可比性 13
 - 12.13 组学专项质量字段 13
- 13 数据安全性与隐私保护 13
 - 13.1 安全管理要求 13
 - 13.2 数据分类分级 13
 - 13.3 处理依据与目的限定 13
 - 13.4 身份认证 13
 - 13.5 最小权限和职责分离 14
 - 13.6 去标识化与假名化 14
 - 13.7 加密与密钥管理 14
 - 13.8 网络和终端安全 14
 - 13.9 日志和审计 14

- 13.10 数据导出 14
- 13.11 第三方和云服务 14
- 13.12 个人权利响应 14
- 13.13 个人信息保护影响评估 14
- 13.14 人类遗传资源信息 14
- 13.15 安全事件响应 14
- 13.16 合规审计 15
- 14 数据应用与服务 15
 - 14.1 数据目录 15
 - 14.2 数据检索 15
 - 14.3 队列构建 15
 - 14.4 临床科研应用 15
 - 14.5 生物标志物研究 15
 - 14.6 模型训练与验证 15
 - 14.7 多中心共享 15
 - 14.8 数据接口服务 15
 - 14.9 数据申请与审批 15
 - 14.10 数据使用监督 16
 - 14.11 成果披露 16
 - 14.12 应用退出 16
- 15 运行维护 16
 - 15.1 运维组织 16
 - 15.2 运行监控 16
 - 15.3 性能管理 16
 - 15.4 容量管理 16
 - 15.5 备份恢复和灾备演练 16
 - 15.6 故障和事件管理 16
 - 15.7 变更管理 16
 - 15.8 发布管理 16
 - 15.9 配置和资产管理 16
 - 15.10 漏洞和补丁管理 17
 - 15.11 供应商管理 17
 - 15.12 文档管理 17
 - 15.13 业务连续性 17
- 16 符合性评价与持续改进 17
 - 16.1 评价原则 17
 - 16.2 建设验收 17
 - 16.3 定期评价 17
 - 16.4 内部审核 17
 - 16.5 管理评审 17
 - 16.6 风险与影响评估 17
 - 16.7 用户反馈 17
 - 16.8 不符合项与纠正措施 18
 - 16.9 持续改进 18

16.10 不符合项分级与判定 18

附 录 A （资料性） 自身免疫病多组学数据库总体架构示意图 19

附 录 B （规范性） 数据库最小数据集及数据分类 20

附 录 C （规范性） 核心元数据和数据字典模板 21

附 录 D （规范性） 患者、样本、检测批次和分析任务编码规则 22

附 录 E （规范性） 数据质量维度、规则和错误代码 23

附 录 F （资料性） 数据访问权限矩阵示例 24

附 录 G （资料性） 数据申请、审批、使用和退出流程 25

附 录 H （资料性） 备份恢复和灾难恢复检查表 26

附 录 I （资料性） 多中心数据映射与接入检查表 27

附 录 J （资料性） 数据血缘和版本记录模板 28

附 录 K （规范性） 自身免疫病核心临床数据集 29

附 录 L （规范性） 不同组学类型最低元数据和质量字段 32

附 录 M （规范性） 符合性评价项目、证实方法和判定规则 33

参考文献 35

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。

本文件由中国欧洲经济技术合作协会提出并归口。

本文件起草单位：。

本文件主要起草人：。

本文件为首次编制。

引言

自身免疫病具有疾病谱广、临床异质性显著、病程长和诊疗数据来源复杂等特点。随着基因组、转录组、表观组、蛋白质组、代谢组、单细胞及空间组学技术的发展，医疗机构和科研单位持续产生规模大、类型多、关联复杂的临床与多组学数据。

多组学数据库不仅是数据存储系统，还应具备统一的数据模型、可验证的数据质量规则、完整的版本和血缘记录、严格的安全与隐私控制以及可持续的运行维护能力。若缺少统一要求，容易出现患者与样本关联错误、临床表型定义不一致、分析结果不可复现、数据共享边界不清和系统安全风险等问题。

本文件面向自身免疫病多组学数据库的建设、治理和应用全过程，重点规定组织治理、总体架构、数据接入、数据标准、预处理与入库、存储和生命周期、版本血缘、数据质量、安全隐私、数据服务及运行维护要求。具体组学检测性能、临床结果解释和诊疗报告由相应专业标准规定。本文件采用“共性核心临床数据、病种扩展数据和组学专项元数据”相结合的组织方式，既覆盖系统性自身免疫病，也覆盖器官特异性自身免疫病；附录所列病种为代表性病种，不构成封闭清单。

自身免疫病多组学数据库建设及应用技术要求

1 范围

本文件规定了自身免疫病多组学数据库的总体原则、组织治理与职责、总体架构、数据来源与接入、数据标准与数据模型、数据预处理与入库、存储与数据生命周期、版本、血缘与可追溯管理、数据质量管理、数据安全与隐私保护、数据应用与服务、运行维护、符合性评价与持续改进。

本文件适用于医疗机构、科研院所、医学检验机构、公共数据平台及经授权的技术服务机构开展自身免疫病多组学数据库的设计、建设、迁移、运行、维护、共享和应用。

本文件适用于单中心和多中心数据库，可覆盖临床数据、样本数据、基因组、转录组、表观组、蛋白质组、代谢组、单细胞组学、空间组学以及其他经验证的数据类型。本文件所称自身免疫病包括系统性自身免疫病和器官特异性自身免疫病，病种范围可涉及风湿免疫、神经、内分泌、消化与肝胆、血液、皮肤等专科领域。附录 K 列出代表性病种的最低扩展数据；未列病种应按照 K.3 建立病种扩展数据集。

按照数据用途，数据库可分为科研型、临床科研型和临床应用型。科研型数据库主要用于机制研究、标志物发现和方法学研究；临床科研型数据库同时承载临床随访和科研分析；临床应用型数据库的数据或模型输出进入诊疗工作流程。临床应用型数据库除符合本文件外，还应符合相应临床技术规范、医疗质量管理和软件验证要求。

本文件不替代组学实验检测技术要求、临床结果解释规则和诊疗报告要求；不适用于未经授权采集、汇聚或公开健康医疗数据的活动。

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 20988 网络安全技术 信息系统灾难恢复规范
- GB/T 22239 信息安全技术 网络安全等级保护基本要求
- GB/T 25070 信息安全技术 网络安全等级保护安全技术要求
- GB/T 28448 信息安全技术 网络安全等级保护测评要求
- GB/T 36073 数据管理能力成熟度评估模型
- GB/T 36344 信息技术 数据质量评价指标
- GB/T 37864 生物样本库质量和能力通用要求
- GB/T 37964 信息安全技术 个人信息去标识化指南
- GB/T 37973 信息安全技术 大数据安全管理指南
- GB/T 38736 人类生物样本保藏伦理要求
- GB/T 39335 信息安全技术 个人信息安全影响评估指南
- GB/T 39725 信息安全技术 健康医疗数据安全指南
- GB/T 39766 人类生物样本库管理规范
- GB/T 39767 人类生物样本管理规范
- GB/T 39768 人类生物样本分类与编码
- GB/T 40364 人类生物样本库基础术语
- GB/T 42460 信息安全技术 个人信息去标识化效果评估指南
- WS/T 303 卫生健康信息数据元标准化规则

- WS/T 305 卫生健康信息数据集元数据标准
- WS/T 306 卫生健康信息数据集分类与编码规则
- WS/T 370 卫生健康信息基本数据集编制标准
- WS/T 671 国家卫生与人口信息数据字典
- WS/T 672 国家卫生与人口信息概念数据模型
- WS/T 787 国家卫生信息资源分类与编码管理规范

3 术语和定义

GB/T 40364 界定的以及下列术语和定义适用于本文件。

3.1

自身免疫病 autoimmune disease

由于免疫耐受失衡，自身反应性免疫细胞、抗体或相关免疫介质异常作用于自身组织和器官，导致慢性炎症、组织损伤或功能障碍的一类疾病。

3.2

多组学 multi-omics

采用两种或两种以上组学技术，对基因组、转录组、表观组、蛋白质组、代谢组或其他分子层级数据进行联合分析，以研究生物系统特征及其相互关系的方法。

3.3

多组学数据库 multi-omics database

对临床信息、样本信息、多组学原始数据、处理后数据、分析结果、模型及相关元数据进行关联、存储、管理和服务的数据库系统。

3.4

源数据 source data

在临床诊疗、样本管理、实验检测或数据管理活动中首次记录事实，并作为后续录入、转换、分析和核验依据的数据。源数据强调事实来源和证据属性，可表现为原始病历记录、样本流转记录、仪器文件或依法形成的业务记录。

3.5

原始数据 raw data

由仪器、设备或信息系统直接产生，尚未经过改变其科学含义的分析处理的数据文件、信号或记录。仅进行文件校验、无损压缩或格式封装，不改变其原始数据属性。

3.6

衍生数据 derived data

由源数据或原始数据经清洗、转换、计算、注释、统计或模型处理后形成的数据。

3.7

元数据 metadata

描述数据的内容、结构、来源、方法、时间、质量、权限、版本和使用限制等特征的数据。

3.8

数据字典 data dictionary

以统一方式描述数据元素名称、定义、类型、格式、值域、单位、约束、来源和校验规则的文件或系统。

3.9

主数据 master data

在多个业务过程或系统中共享并保持相对稳定，用于统一识别核心实体的数据。

3.10

患者主索引 enterprise master patient index; EMPI

在授权范围内对同一患者在不同系统、机构或时间点的身份记录进行匹配和统一标识的机制。

3.11

样本唯一标识 unique specimen identifier

在样本全生命周期内唯一识别样本及其分装、衍生和流转关系的编码。

3.12

数据血缘 data lineage

描述数据从来源、采集、转换、分析到输出的生成路径及其依赖关系的信息。

3.13

数据溯源 data provenance

用于证明数据来源、形成过程、责任主体、操作记录和真实性的证据信息。

3.14

数据版本 data version

用于识别数据集在特定时间点的内容、结构和状态的唯一标记。

3.15

数据质量规则 data quality rule

用于判断数据是否满足完整性、准确性、一致性、唯一性、及时性、有效性和可追溯性要求的可执行规则。

3.16

去标识化 de-identification

通过对个人信息进行技术处理，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。

3.17

假名化 pseudonymization

使用非直接身份标识替代真实身份标识，并将映射关系分离保管的处理方式。

3.18

数据共享 data sharing

在明确目的、授权范围和安全控制下，向其他主体提供数据访问、查询、计算或传输服务的活动。

3.19

受控分析环境 controlled analysis environment

对身份、权限、数据导入导出、计算过程和操作日志实施控制的数据分析环境。

3.20

数据冻结 data freeze

将经审核的数据集固定为不可随意修改的版本，用于分析、验证、发布或归档的过程。

3.21

数据归档 data archiving

将不再频繁使用但需继续保留的数据及其必要元数据迁移至受控保存环境的过程。

3.22

数据销毁 data destruction

采用经验证的方法使数据及其副本无法恢复或再次使用的过程。

3.23

匿名化 anonymization

对个人信息进行处理，使其无法识别特定自然人且不能复原的过程。匿名化应与去标识化、假名化相区分，并在公开、共享或成果披露前评估重识别风险。

4 总体原则**4.1 依法合规**

数据库建设和应用应符合医疗卫生、网络安全、数据安全、个人信息保护、生物安全、医学伦理和人类遗传资源管理等要求。数据用途、处理依据、责任主体和授权范围应在处理前明确。

4.2 临床与科研需求驱动

数据库应服务于明确的临床、科研、质量管理或公共卫生目标。不得以“先汇聚、后确定用途”为由无限制收集数据，新增数据类型和应用场景应开展必要性 with 风险评估。数据库类型、预期用途、允许输出和禁止用途应纳入数据目录并保持版本记录。

4.3 统一标准

患者、样本、疾病、时间点、检测批次、分析流程和结果对象应采用统一的数据元素、编码、术语、格式、单位和交换规则。无法直接采用统一标准时，应建立可追溯的映射关系。

4.4 最小必要

采集、存储、使用、共享和保留的数据应与处理目的直接相关，并限定至实现目的所需的最小范围、最短期限和最低权限。

4.5 分层管理

应根据数据类型、敏感程度、可识别性、使用目的和泄露后影响实施分类分级，并据此配置访问、传输、存储、共享和销毁措施。

4.6 安全内生

安全和隐私保护应嵌入需求、架构、开发、测试、部署、运行、变更和退出全过程，不得在系统建成后仅通过补充外围控制代替安全设计。

4.7 全程可追溯

数据来源、采集、转换、版本、质量状态、访问、修改、导出、共享、分析和删除应可追溯。关键衍生结果应能够追溯至源数据、流程和参数。

4.8 原始数据保护

原始数据应与衍生数据分离管理，未经授权不得覆盖或改写。更正应保留原记录、更正原因、责任人和时间。

4.9 可扩展与互操作

数据库架构应支持新增疾病、中心、组学类型、分析流程和数据服务，并通过标准接口和共同数据模型提高跨系统、跨中心互操作能力。

4.10 质量可量化

数据质量应通过可执行规则、指标和报告进行评价。质量状态应随数据流转，并在数据检索、导出和应用时可见。质量阈值应按科研、临床科研和临床应用等用途分级制定，不得以同一阈值覆盖风险明显不同的场景。

4.11 责任可落实

数据来源单位、数据库管理单位、技术服务单位和数据使用者的职责应通过制度和协议明确，责任不得因委托处理或云服务而转移或消失。

4.12 持续改进

应基于质量问题、安全事件、用户反馈、技术变更和合规要求持续评价并改进数据库。

5 组织治理与职责

5.1 治理体系

数据库建设单位应建立覆盖决策、数据治理、质量、安全、伦理合规、技术运维和用户服务的治理体系。治理体系应与数据库规模、数据敏感性、应用风险和多中心协作复杂度相适应。

宜设立数据库治理委员会或具有同等职能的组织，对数据范围、标准、重大变更、共享应用、风险和资源投入进行决策。治理组织宜包含风湿免疫或相关专科医师、实验检测人员、生物信息人员、数据治理人员、信息安全人员和伦理合规人员。

5.2 角色与职责

要求如下：

- a) 数据主管负责数据政策、数据标准、数据资产目录和跨部门协调；
- b) 数据责任人负责特定数据域的定义、来源合法性、质量目标和授权范围；
- c) 数据管理员负责数据字典、元数据、质量规则、问题处置和版本维护；
- d) 系统负责人负责架构、开发、部署、性能、备份、恢复和变更管理；
- e) 安全负责人负责分类分级、访问控制、日志审计、风险评估和安全事件响应；
- f) 伦理与合规负责人负责知情同意、二次利用、共享、对外提供和人类遗传资源相关审查；
- g) 数据使用者应按批准目的和范围使用数据，并对导出数据、分析环境和成果披露承担责任。

5.3 数据来源单位责任

数据来源单位应保证提供数据的来源合法、标识准确、内容真实、格式符合协议，并及时通报更正、撤回、质量问题和授权范围变化。来源单位不得将无法证明来源、未经授权或已知存在重大质量缺陷的数据接入正式库。

5.4 委托处理与共同处理

5.4.1 委托技术服务、云服务或联合建设时，应在合同或协议中明确处理目的、数据范围、处理方式、保密义务、安全措施、分包限制、审计权、事件通报、服务终止、数据返还和销毁要求。

5.4.2 共同决定处理目的和方式的多方应明确各自责任、个人权利响应机制和对外联络渠道。

5.5 制度与文件

至少应建立数据接入、数据标准、质量管理、访问授权、数据共享、备份恢复、变更发布、安全事件、个人权利响应、第三方管理和数据销毁制度。制度应定期评审并保持有效版本。

5.6 培训与授权

人员应按岗位接受数据治理、质量、安全、隐私和伦理培训。关键岗位应通过能力确认后授权，人员调岗或离职时应及时调整或撤销权限并完成数据、账号和文档交接。涉及临床数据定义、疾病活动度量表、样本前处理和组学数据管理的人员，还应接受相应专业术语、量表版本和组学最低元数据要求培训。

5.7 利益冲突与监督

涉及数据审批、质量判定、安全审计和重大变更的人员应避免利益冲突。必要时应引入独立复核或第三方评价。

6 总体架构

6.1 架构组成

数据库总体架构应至少包括业务架构、数据架构、应用架构、技术架构和安全架构。各架构之间应有明确接口，避免临床系统、样本系统、组学平台和分析平台形成不可关联的数据孤岛。

6.2 业务架构

业务架构应覆盖数据产生、采集、接入、标准化、质控、入库、查询、分析、共享、归档、删除和销毁。每一业务环节应明确输入、输出、责任人、审批节点和异常处理。

6.3 数据架构

数据架构应将临床、样本、组学原始数据、处理后数据、分析结果、模型、元数据、质量、权限和审计数据分层管理，并建立患者—就诊—疾病事件—治疗事件—样本—分装—检测批次—分析任务—结果—随访结局的关联链。

6.4 应用架构

应用架构宜包括数据目录、数据接入、主数据、元数据、数据质量、样本关联、队列构建、检索、分析、共享审批、审计和运维监控等组件。应用组件应通过受控接口访问数据，不宜直接绕过服务层操作生产数据库。

6.5 技术架构

技术架构应根据数据规模和访问模式配置计算、存储、网络、数据库、中间件、流程管理、容器和监控资源。应支持结构化、半结构化和非结构化数据，并对大文件、稀疏矩阵、影像和组学对象提供适配的存储与索引。

6.6 安全架构

安全架构应覆盖身份认证、权限控制、网络隔离、传输与存储加密、密钥管理、恶意代码防护、漏洞管理、日志审计、数据防泄漏、备份恢复和安全监测。网络安全等级保护应符合 GB/T 22239、GB/T 25070 和 GB/T 28448 的相关要求。

6.7 部署方式

可采用本地部署、专有云或合规的混合部署。选择部署方式时应评估数据敏感性、数据所在地、服务连续性、网络条件、供应商锁定、退出迁移和监管要求。未经评估，不得将含直接标识符或高敏感数据部署于无法实施有效控制的公共环境。

6.8 环境隔离

生产、测试、开发和培训环境应隔离。生产数据原则上不得直接复制到开发或测试环境；确需使用时，应经过审批、最小化和去标识化，并在使用结束后销毁。

6.9 高可用与扩展

关键服务应根据业务影响设置高可用、容量冗余和故障转移措施。架构应支持水平或垂直扩展，并避免单点故障导致关键数据不可用或不可恢复。

6.10 架构文档

应形成总体架构图、数据流图、网络拓扑、接口清单、资产清单和信任边界说明，并在重大变更后及时更新。总体架构示例见附录 A。

7 数据来源与接入

7.1 数据来源登记

每一数据来源应登记来源单位、系统或设备、责任人、采集目的、处理依据、数据类型、覆盖时间、更新频率、授权范围、质量状态和退出条件。外部公开数据还应记录许可条款、引用方式和使用限制。

7.2 临床数据

7.2.1 临床数据应采用“共性核心数据+病种扩展数据”的两层结构。共性核心数据至少应覆盖人口学、发病与确诊时间、诊断状态、诊断标准及版本、疾病亚型、疾病活动度、器官受累与累积损伤、自身抗体谱、补体和免疫球蛋白、炎症与常规实验室指标、影像学、治疗暴露、疗效、不良事件、合并症、特殊临床事件、随访和结局。病种扩展数据应根据类风湿关节炎、系统性红斑狼疮、干燥综合征、系统性硬化症、炎性肌病、抗磷脂综合征、系统性血管炎、1 型糖尿病、自身免疫性甲状腺病、多发性硬化、重症肌无力、炎症性肠病、乳糜泻、自身免疫性肝病、自身免疫性血液病和自身免疫性皮肤病等疾病特点设置。核心数据见附录 K。

7.2.2 疾病活动度、器官损伤、缓解、低疾病活动、复发或治疗反应等复合指标，不应仅保存总分或最终分级；在数据可获得时，应同时保存量表名称、量表版本、评价时间、构成项目、缺失项目、总分、判定结果和评价人员。

7.2.3 自身抗体、补体、免疫球蛋白和炎症指标应保留原始结果、检测方法、单位、参考范围、阳性界值、实验室和检测时间。药物数据应区分糖皮质激素、传统改善病情抗风湿药、免疫抑制剂、生物制剂和靶向合成药物，并记录剂量、途径、开始和结束时间、停药原因、疗效和不良事件。

7.3 样本数据

样本数据应覆盖采集、容器、处理、分装、保存、位置、冻融、质量、借用、归还、耗用、销毁和知情同意限制，并记录采样时疾病状态、治疗状态、感染情况、空腹状态、采样时间及其他可能影响组学结果的前分析因素。样本管理应符合 GB/T 37864、GB/T 39766、GB/T 39767 和 GB/T 39768 的相关要求；人类生物样本保藏和二次利用的伦理管理应符合 GB/T 38736 的相关要求。知情同意、伦理批准、二次利用、共享和撤回状态应以结构化字段管理，并与样本及其衍生数据关联。

7.4 多组学原始数据

原始数据应记录文件名称、格式、大小、校验值、生成设备、软件、运行编号、批次、样本映射、质控结果和生成时间。原始文件接入前应完成完整性和恶意代码检查。不同组学类型还应保存相应的实验设计、前处理、平台、参考资源和关键质量字段，最低要求见附录 L。

7.5 处理后和分析数据

处理后数据可包括表达矩阵、变异文件、峰值、蛋白或代谢物定量、细胞注释、通路结果和模型输出。每一结果应关联源数据、分析流程、参数、参考资源和质量状态。不得仅保存可视化图片或不可计算的汇总结果；用于复现和再分析的机器可读矩阵、特征文件、注释表和质量报告应同步保存。

7.6 外部数据

接入公共数据库、知识库、指南、文献或商业数据时，应核验来源权威性、版本、许可、更新周期和撤回机制。外部数据不得与本地数据混淆为自行产生的数据。

7.7 接入协议

接入前应形成数据接入协议，至少规定数据集名称、字段、格式、单位、编码、频率、传输方式、校验规则、质量阈值、错误反馈、重传、版本、责任和停用条件。

7.8 安全传输

敏感数据传输应采用身份认证和加密措施，并对传输端、接收端、文件校验值、时间和状态进行记录。使用移动介质时，应经过批准、加密和病毒检查，并登记介质流转。

7.9 身份映射

应建立患者、就诊、样本、检测批次和分析任务之间的可靠映射。直接标识符与研究标识符应分离管理，映射表应单独保护并限制访问。家系成员应通过独立的家系关系表关联，不应在患者研究标识中直接编码亲属关系或可识别身份信息。

7.10 接入验收

首次接入和重大变更后应进行验收，至少验证字段、格式、编码、数量、时间范围、完整性、重复、关联、权限和性能。未通过验收的数据不得进入正式库。

7.11 增量更新与撤回

应明确全量、增量和更正数据的更新方式。来源数据被更正或撤回时，应识别并处理受影响的衍生数据、分析结果和已导出数据。

8 数据标准与数据模型

8.1 数据分类与分层

数据至少应分为临床数据、样本数据、组学原始数据、处理后数据、分析结果、模型数据、元数据、质量数据、权限数据和审计数据。不同层级应明确是否可修改、保留期限和访问范围。

8.2 患者主索引

患者主索引应支持同一患者多来源记录的匹配、冲突识别、人工复核、合并和拆分。自动匹配规则应经过验证，置信度不足时不得自动合并。每次合并或拆分应保留审计记录。

8.3 样本标识与关系

样本编码应在数据库范围内唯一，并记录母样本、分装、衍生样本和耗用关系。样本唯一标识不宜直接包含姓名、身份证件号码等直接标识信息。编码规则见附录 D。

8.4 临床数据模型

临床数据模型应支持诊断事件、疾病活动度、器官受累、治疗方案、用药时间、剂量、疗效、不良事件、随访和结局。应区分当前状态、历史状态、疑似诊断和确诊诊断。诊断标准、量表和结局定义应作为可版本化对象管理；同一疾病采用不同标准或量表时，应保留原始项目并建立映射，不得仅转换为统一总分。

8.5 多组学数据模型

多组学数据模型应至少描述组学类型、实验设计、样本、批次、平台、原始文件、特征、注释、质控、分析任务、结果对象和参考资源。不同组学特征应使用稳定标识符，并记录标识符系统和版本。各组学类型应按照附录 L 建立最低元数据和质量字段，并区分实验质控、数据质控和结果质控。

8.6 元数据模型

元数据应覆盖业务、技术、操作、质量、安全和管理信息。核心元数据不得依赖文件名或人工记忆，应以结构化方式保存。模板见附录 C。卫生健康数据集元数据应符合 WS/T 305 的相关要求。

8.7 数据字典

数据字典应包括字段中文名称、英文名称、定义、数据类型、长度、格式、值域、单位、必填性、可空含义、来源、转换规则、质量规则、敏感级别和版本。字段含义变化时应创建新版本，不得仅修改说明而不留痕。数据元命名、定义和表示格式应符合 WS/T 303、WS/T 370 和 WS/T 671 的相关要求。

8.8 术语和代码集

疾病、药品、检查、样本和结果等应优先采用适用的国家、行业或公认医学术语和代码集。使用本地代码时应建立映射，并记录一对多、多对一和无法映射的情况。数据集分类、资源编码和跨系统映射应符合 WS/T 306、WS/T 672 和 WS/T 787 的相关要求。

8.9 单位和精度

同类指标应使用统一单位和允许精度。单位换算应保留原始值、原始单位、转换值、目标单位、换算规则和执行版本。

8.10 时间与事件模型

应统一日期时间格式、时区和精度，并区分采集时间、入库时间、事件发生时间、报告时间和更新时间。基线、治疗、复诊、复发和结局等事件应通过时间关系进行关联。

8.11 缺失和未知状态

应区分未采集、不适用、未知、拒绝提供、技术失败、低于检出限和数据丢失等状态，不得使用同一空值表示不同含义。

8.12 完整性约束

应定义主键、外键、唯一性、非空、值域、时间顺序和跨表关系约束。孤立样本、无数据溯源结果和无版本分析任务应被识别并限制使用。

8.13 模型变更管理

数据模型、字段、代码集和接口的变更应评估对既有数据、应用、接口和分析结果的影响。重大变更应提供迁移方案、兼容期和回退方案。

9 数据预处理与入库

9.1 总体流程

数据入库应按“接收—暂存—校验—清洗—转换—质控—审核—正式入库”的流程执行。处理步骤应自动记录并支持复现。

9.2 暂存区

新接入数据在通过验收前应存放于与正式库隔离的暂存区。暂存区应限制访问、设置保留期限并防止数据被误用于分析。

9.3 文件与语法校验

应检查文件格式、编码、压缩包结构、校验值、字段数量、字段类型、日期格式、必填项和文件完整性。损坏、未完成传输或格式未知的文件应隔离。

9.4 数据清洗

清洗应依据预先批准的规则处理空格、大小写、格式差异、非法字符、单位、术语和明显错误。不得通过自动清洗改变原始临床含义。自动修正应保留修正前后值和规则。

9.5 重复和冲突

应建立记录级和实体级重复识别规则。冲突处理应依据来源优先级、时间、签名状态和人工复核，不得无证据覆盖。

9.6 异常值

异常值应根据业务范围、统计分布、设备范围和跨字段关系识别。异常值不得仅因偏离总体分布而删除，应区分真实极端值、录入错误和技术失败。

9.7 组学预处理结果

组学预处理应记录软件、参数、参考资源、批次、过滤标准、缺失值处理和质量指标。数据库可接收多种分析流程产生的结果，但不得将不同版本结果合并为无版本标识的单一结果。组学结果接入时应验证附录 L 所列关键元数据和质量字段是否完整；缺失关键字段的数据应标记为限制使用或待复核。

9.8 质量状态标识

数据应至少标识为“通过”“警告”“失败”“待复核”或“撤回”等状态，并记录规则、原因、时间和责任人。质量状态应在查询、导出和分析时同步提供。

9.9 入库审核

正式入库前应完成自动校验和必要的人工复核。高风险数据、关键关联和批量更正应实行双人复核或等效控制。

9.10 入库日志

应记录数据来源、数据集版本、文件校验值、处理步骤、规则版本、记录数量、通过和失败数量、操作者、审核者及时间。

9.11 回滚与重新处理

入库流程应支持按数据批次回滚。重新处理时应创建新版本，保留原版本并明确替代关系。

9.12 数据冻结与发布

用于正式分析、模型验证或对外共享的数据集应冻结并分配唯一版本。冻结后不得直接修改；更正应通过新版本发布。

10 存储与数据生命周期

10.1 分层存储

应根据访问频率、数据规模、敏感等级、性能需求和保留期限设置在线、近线、归档等存储层。数据迁移不得导致文件校验值、元数据和访问控制丢失。

10.2 原始与衍生数据分离

原始数据应以只读或等效方式保护。衍生数据应通过血缘关系关联原始数据，不得覆盖原始文件。

10.3 容量管理

应监测存储使用率、增长率、文件数量、对象大小和性能，设置预警阈值并制定扩容计划。扩容不得降低安全、完整性和备份能力。

10.4 数据完整性

大文件和关键数据应保存校验值，并在传输、迁移、备份和恢复后验证。发现校验不一致时应隔离并调查。

10.5 备份

备份范围应包括数据库、对象存储、配置、密钥材料的可恢复副本、数据字典、元数据、代码、流程和必要文档。备份频率和保留周期应根据业务影响和恢复目标确定。至少应有一份备份与生产环境物理或逻辑隔离。

10.6 恢复验证

应定期进行恢复测试，验证数据可读性、完整性、权限、关联和应用功能。仅完成备份而未验证恢复能力，不视为满足要求。

10.7 灾难恢复

应依据 GB/T 20988 确定灾难恢复等级、恢复时间目标和恢复点目标，形成灾难恢复预案并开展演练。

10.8 保存期限

保存期限应综合法律法规、伦理同意、临床科研用途、合同、数据价值和风险确定。期限届满前应复核继续保存的必要性。

10.9 归档

归档数据应保留必要的索引、格式说明、读取工具或转换方案，确保长期可读。数据归档操作应记录数据范围、版本、介质、位置和校验结果。

10.10 删除与销毁

删除和销毁应经过审批。生产环境、缓存、临时文件和可控导出副本应按批准要求及时删除或销毁。对于不可变备份、增量备份或灾难恢复介质中无法即时选择性删除的数据，应建立删除清单和恢复后再删除机制，限制相关数据恢复至正常业务使用，并在备份到期时完成销毁。采用逻辑删除时应明确适用范围、屏蔽措施和最终销毁时间。销毁应形成记录并验证不可恢复或不可继续使用。

10.11 服务退出和数据迁移

更换系统、供应商或终止服务时，应制定数据迁移、校验、业务切换、旧系统封存和数据销毁方案，防止因格式锁定或供应商退出导致数据不可用。

11 版本、血缘与可追溯管理

11.1 数据集版本

数据集发布、修订、冻结、撤回和归档应使用唯一版本号。版本说明应列明时间范围、数据来源、记录数量、数据结构、质量状态和与前一版本的差异。

11.2 参考资源版本

应记录参考基因组、注释库、通路库、蛋白质库、代谢物库、医学术语和知识库的名称、版本、发布日期或访问日期。参考资源更新后，应评估是否需要重新分析。

11.3 软件与流程版本

应记录软件、算法、参数、工作流、容器、操作系统、运行环境和依赖。流程版本应经过测试和批准后发布。

11.4 模型版本

模型应记录训练和验证数据版本、特征、算法、超参数、阈值、性能、适用范围、限制、负责人和发布日期。不得用相同名称覆盖不同模型。本文件规定模型在数据库中的登记、版本和追溯要求；模型的临床性能、临床阈值和报告解释应由相应临床技术规范规定。

11.5 数据血缘

血缘应覆盖源系统、接入批次、清洗规则、转换、分析流程、中间产物和输出。关键结果应支持从结果反向定位至源数据，也应支持从源数据识别受影响的下游结果。数据血缘和版本记录模板见附录 J。

11.6 人工操作留痕

人工更正、合并、拆分、标注和结果确认应记录原值、新值、原因、证据、操作者、审核者和时间。不得使用共享账号完成关键操作。

11.7 审计追踪

应记录新增、修改、删除、查询、导出、共享、权限变更、登录失败和管理操作。日志应防篡改，并与业务数据分离保存。

11.8 更正、撤回和影响分析

发现源数据错误、授权撤回或分析流程缺陷时，应评估受影响的数据集、分析、模型和共享对象，并通知相关使用者。更正不得删除原证据。

11.9 可复现性验证

对重要分析流程和发布数据集，应抽样验证在相同输入、版本和环境下能否复现主要输出。无法复现时应查明原因并限制使用。

12 数据质量管理

12.1 质量管理体系

应建立覆盖数据全生命周期的质量管理体系，明确质量目标、规则、责任、监测、问题处置、报告和改进。数据质量评价应符合 GB/T 36344 的相关要求。

12.2 质量维度

质量评价至少应包括完整性、准确性、一致性、唯一性、及时性、有效性和可追溯性。不同数据类型和用途可增加可比性、稳定性、可复现性和代表性等维度。

12.3 质量目标和阈值

应根据临床、科研、模型训练和共享等用途分别设定质量目标。阈值应有业务和技术依据，并经批准；不得将探索性阈值作为所有用途的统一合格标准。

12.4 质量规则

质量规则应可执行、可版本化并有唯一编号，至少包括字段级、记录级、跨表、跨时间、跨中心和组学专项规则。规则示例见附录 E。

12.5 来源质量

应监测来源系统的缺失、延迟、映射失败和异常变化。数据质量问题应尽可能在来源端纠正，不宜长期依赖数据库端补丁。

12.6 样本与组学质量关联

样本质量、实验质控和数据质控应与结果关联。质量不合格或关键质控缺失的数据不得与合格数据无差别提供。

12.7 运行质量监测

应持续监测接口失败、数据量突变、字段分布漂移、编码变化、关联丢失、重复增加和质量趋势。发现异常时应确认是否为真实业务变化。

12.8 问题分级

质量问题应根据影响范围、临床或科研风险、可恢复性和紧急程度分级，并规定响应时限、责任人和关闭条件。

12.9 不合格数据

不合格数据应隔离或明确标识，并根据风险采取限制使用、修复、重新接入、撤回或作废。不得在未说明质量状态的情况下对外提供。

12.10 质量报告

应定期形成质量报告，至少包括数据规模、规则执行、问题分布、趋势、重大问题、整改和未解决风险。质量结果应提供给治理组织和数据使用者。

12.11 质量抽查和复核

应对自动规则无法覆盖的语义准确性、临床逻辑和复杂关联开展人工抽查。抽样方案、结果和纠正措施应记录。核心临床字段、患者一样本关联和高风险数据至少每年开展一次专项抽查；抽样范围应覆盖不同中心、时间段和数据来源。

12.12 多中心可比性

多中心数据库应比较中心间字段、单位、疾病定义、样本流程、平台和数据分布。不可比差异应通过映射、分层分析或限制说明处理，不得简单合并。应识别同一患者跨中心重复、同一患者多时间点样本和家系相关个体；用于统计分析或模型训练时，应按研究设计进行患者级或家系级分组，避免将相关记录误作为独立样本。

12.13 组学专项质量字段

数据库应按照附录 L 对基因组、转录组、表观组、单细胞组学、空间组学、蛋白质组、代谢组等数据建立最低元数据和质量字段。未达到既定质量阈值的数据可保留用于方法学研究，但应显著标识质量状态、限制用途和不确定性。

13 数据安全与隐私保护

13.1 安全管理要求

数据安全应符合 GB/T 39725 和 GB/T 37973 的相关要求。应建立数据安全责任制、风险评估、监测预警、事件响应和持续改进机制。

13.2 数据分类分级

应按数据类型、敏感程度、可识别性、规模、用途和泄露后影响实施分类分级。基因组、健康状况、身份标识和可重识别的多组学组合数据应作为高敏感数据重点保护。

13.3 处理依据与目的限定

处理个人信息应具有明确、合理的目的和适用的处理依据。目的、范围或共享对象发生实质变化时，应重新开展合规和必要性评估，并根据要求履行告知、同意或伦理程序。

13.4 身份认证

用户应使用唯一账号。高权限、远程访问和敏感数据导出宜采用多因素认证。账号生命周期应覆盖申请、批准、开通、复核、调整、停用和注销。

13.5 最小权限和职责分离

应基于角色和必要属性授权，并按数据域、功能、用途和时间限制权限。数据审批、系统管理、安全审计和业务使用等关键职责应适当分离。权限矩阵示例见附录 F。

13.6 去标识化与假名化

13.6.1 用于科研、共享、测试和模型训练的数据应根据目的实施去标识化或假名化。处理方法和效果评价应符合 GB/T 37964 和 GB/T 42460 的相关要求。需要公开发布的数据应优先采用匿名化、统计汇总或受控访问方式；不得将去标识化等同于匿名化。

13.6.2 直接标识符、映射表和研究数据应分离存储。重识别仅可在批准的临床核验、个人权利响应或法律要求等场景下进行，并记录理由和操作。

13.7 加密与密钥管理

敏感数据在传输、存储和备份中应采用适当的加密或等效保护。密钥应与数据分离管理，明确生成、分发、使用、轮换、备份、吊销和销毁要求。

13.8 网络和终端安全

应实施网络区域划分、边界控制、漏洞修复、恶意代码防护、终端管控和远程访问管理。生产数据库不应直接暴露于互联网。

13.9 日志和审计

应记录对敏感数据的查询、浏览、下载、导出、修改、共享和删除。日志应包含用户、时间、对象、操作、来源地址和结果，并采用防篡改措施。日志保留期限应满足法律法规、等级保护和业务追溯要求。

13.10 数据导出

导出应经过申请和批准，限定字段、记录、格式、用途、接收人和有效期。高敏感数据宜通过受控分析环境使用，减少原始数据离开安全边界。对稀有疾病、极端表型、家系关系和罕见基因变异等可能导致重识别的输出，应设置最小输出单元、聚合、抑制或人工审核。

13.11 第三方和云服务

使用第三方或云服务前应开展安全和隐私评估，明确数据位置、访问权限、服务水平、漏洞和事件通报、审计、分包、退出和销毁。供应商不得利用数据训练未获授权的模型或开展其他用途。

13.12 个人权利响应

应建立查询、复制、更正、删除、撤回同意和解释等请求的受理、身份核验、评估和答复流程。涉及研究数据和法定保留义务时，应说明适用限制。

13.13 个人信息保护影响评估

处理敏感个人信息、开展自动化决策、对外提供、跨境提供或实施其他高风险活动前，应按照 GB/T 39335 及适用法规开展个人信息保护影响评估，并保存记录。

13.14 人类遗传资源信息

涉及人类遗传资源信息的采集、保藏、利用、对外提供和国际合作，应设置专项合规审查。应区分临床诊疗所需的数据处理与科研性采集、保藏、利用和共享活动；属于临床诊疗范围的，按照医疗卫生相关规定执行，涉及科学研究、对外提供或国际合作的，应按照现行人类遗传资源管理要求履行相应程序。不得以数据库建设、技术服务或模型训练名义规避适用要求。

13.15 安全事件响应

应建立发现、报告、隔离、调查、取证、通知、恢复和改进流程。事件发生后应评估数据类型、数量、影响对象、可重识别性和下游共享范围，并按要求报告和通知。

13.16 合规审计

应定期审查个人信息处理、权限、导出、共享、第三方和安全措施的合规性。审计发现应形成整改计划并验证效果。

14 数据应用与服务

14.1 数据目录

应向授权用户提供数据目录，说明数据集名称、范围、字段、时间、质量、版本、敏感级别、更新频率、可用方式和限制。目录不得暴露未授权的个人信息。

14.2 数据检索

检索应基于授权范围，并记录查询条件和结果规模。对可能通过小样本组合识别个人的查询，应设置最小输出、聚合或人工审核。

14.3 队列构建

队列定义应结构化记录纳入、排除、时间窗、疾病定义、数据版本和执行时间，以支持重复构建。队列人数和特征变化时应能解释其原因。队列定义还应记录索引事件、基线窗口、暴露窗口、随访起点、结局窗口和删失规则。

14.4 临床科研应用

科研应用应基于批准的项目、伦理和数据使用协议。应遵循目的限定、最小数据集和期限控制，研究结束后按协议返还、归档或销毁数据。数据库最小数据集见附录 B。

14.5 生物标志物研究

数据库可支持标志物发现、验证和复现，但应区分发现数据集、验证数据集和独立测试数据集。研究性结果不得直接作为临床诊断或治疗结论。

14.6 模型训练与验证

14.6.1 模型训练应记录数据版本、授权、样本选择、标签、特征、拆分策略、缺失处理和偏倚控制。应防止同一患者或相关样本跨训练和测试集合造成数据泄漏。存在家系关系时，应根据模型用途采用家系级拆分；多中心模型宜采用中心外验证、时间外验证或独立外部验证，并记录中心、平台和人群差异。

14.6.2 对外提供模型或模型输出时，应说明适用人群、性能、限制、版本和监测要求。不得以模型输出反向推断未授权个人信息。

14.7 多中心共享

多中心共享应采用共同数据模型、映射规则和质量评价。可根据风险采用受控传输、可信研究环境、联邦分析或安全多方计算等方式。共享前应完成同一患者跨中心重复识别、病种定义对齐和组学平台差异评估。多中心数据映射与接入检查见附录 I。

14.8 数据接口服务

接口应实施认证、授权、传输加密、限流、输入校验、错误处理、日志和版本管理。接口停用或升级应提前通知并提供迁移方案。

14.9 数据申请与审批

数据申请应说明目的、法律或伦理依据、数据范围、分析方法、人员、环境、期限、输出和销毁计划。审批应综合评价必要性、质量适用性、安全风险和利益冲突。流程见附录 G。

14.10 数据使用监督

应通过日志、进度报告、成果审查或现场检查等方式监督使用情况。发现超范围使用、未授权复制或安全措施失效时，应暂停访问并调查。

14.11 成果披露

论文、报告、图表和公开数据应经过披露风险审查，避免通过稀有疾病、家系、极端表型、基因变异或多变量组合识别个人。

14.12 应用退出

项目到期、人员离开、授权撤回或用途终止时，应关闭权限，核验数据返还或销毁，回收移动介质，并形成退出记录。

15 运行维护

15.1 运维组织

应明确系统、数据库、存储、网络、安全、数据质量和用户支持岗位，建立值守、升级和应急联系机制。关键岗位应有替代人员。

15.2 运行监控

应监测服务可用性、接口、作业、数据库、存储、计算、备份、日志和安全状态。监控告警应分级、通知、处置和关闭。

15.3 性能管理

应定期评价查询、入库、批处理、文件传输和分析作业性能，识别资源瓶颈。性能优化不得绕过权限、审计或完整性控制。

15.4 容量管理

应根据数据增长、并发、备份和归档需求预测容量。达到预警阈值前应完成扩容或迁移。

15.5 备份恢复和灾备演练

应按计划执行备份、恢复测试和灾难恢复演练。演练应验证人员、通信、数据、系统和业务流程，并形成问题和整改记录。检查表见附录 H。数据恢复验证至少每年开展一次；临床应用型数据库至少每年开展一次灾难恢复演练，科研型和临床科研型数据库的演练周期原则上不应超过两年。

15.6 故障和事件管理

故障应按影响分级，记录发现、响应、升级、临时措施、恢复、根因和预防措施。重复故障应纳入问题管理。

15.7 变更管理

数据库结构、接口、软件、硬件、网络、规则和配置变更应经过申请、影响评估、测试、审批、发布、验证和回退准备。紧急变更应在事后补充审核。患者匹配规则、核心数据模型、组学处理流程、权限模型、平台迁移和临床模型版本变化应作为重大变更管理。

15.8 发布管理

发布包应包含版本、内容、依赖、测试结果、变更说明、部署步骤、回退方案和负责人。生产发布应由授权人员执行并记录。

15.9 配置和资产管理

应维护服务器、存储、网络、数据库、软件、账号、证书、许可证和配置基线。未知或未授权资产不得接入生产环境。

15.10 漏洞和补丁管理

应识别和评估漏洞，按风险制定修复时限。补丁应在测试环境验证后部署；无法及时修复时应采取临时缓解措施。

15.11 供应商管理

应评价供应商能力、稳定性、安全和合规表现，明确服务水平、审计权、人员保密、远程运维、事件通报和退出交接。

15.12 文档管理

架构、数据字典、接口、流程、运维、安全、应急和用户文档应受控管理。过期文档应标识或撤回，确保现场使用有效版本。

15.13 业务连续性

应识别关键业务和最大可容忍中断时间，制定替代方案、恢复优先级和沟通计划。人员、场地、供应链和网络中断均应纳入考虑。

16 符合性评价与持续改进

16.1 评价原则

符合性评价应基于文件审查、现场检查、配置核验、数据抽样、技术测试、日志检查、访谈和演练等证据，不应仅依据制度文本或自我声明。

16.2 建设验收

数据库首次上线、重大迁移或架构重建前，应对功能、数据、质量、安全、性能、备份恢复、接口和文档进行验收。未关闭的高风险问题应经治理组织评估并明确限制措施。验收项目、证实方法和判定规则见附录 M。存在未关闭的关键不符合项或重大不符合项时，不应通过验收。

16.3 定期评价

应按计划评价治理、数据质量、安全、性能、合规、用户服务和业务价值。评价频率应与风险和变更程度相适应。定期评价原则上每年至少开展一次，可参考 GB/T 36073 评价数据管理能力成熟度；发生重大安全事件、重大质量问题或重大变更后应开展专项评价。

16.4 内部审核

内部审核应由与被审核活动相对独立的人员实施，覆盖制度执行、数据处理记录、权限、共享、运维和整改。内部审核原则上每年至少开展一次，高敏感数据大量导出、跨境共享或临床应用型数据库可根据风险增加频次。

16.5 管理评审

管理评审应考虑质量和安全指标、事件、用户反馈、资源、重大变更、法规和标准变化，形成决策、责任人和完成时限。

16.6 风险与影响评估

新增数据类型、新用途、新共享对象、跨境场景、云迁移、自动化决策和重大模型应用前，应开展风险、数据安全和个人信息保护影响评估。

16.7 用户反馈

应建立数据错误、字段需求、接口故障、结果不一致和服务建议的反馈渠道。反馈应分类、响应、关闭并用于改进。

16.8 不符合项与纠正措施

不符合项应记录事实、范围、影响和临时控制，开展根因分析，制定纠正措施并验证效果。重大问题应评估是否需要通知数据使用者或撤回数据。

16.9 持续改进

应形成“问题识别—风险评估—措施制定—实施—效果验证—标准和制度更新”的闭环。改进不得以降低安全或数据可追溯性为代价。

16.10 不符合项分级与判定

16.10.1 不符合项应分为关键不符合项、重大不符合项、一般不符合项和观察项。关键不符合项是可能导致患者、样本或结果错误关联，造成严重数据泄露、不可恢复丢失或临床安全风险的问题；重大不符合项是对质量、安全、合规或可追溯性造成显著影响的问题；一般不符合项是局部偏差且风险可控的问题；观察项是尚未构成不符合但存在改进空间的问题。

16.10.2 符合性评价通过应同时满足：无关键不符合项，无未关闭的重大不符合项；一般不符合项已完成整改或形成经批准的整改计划；抽样、测试和演练证据能够支持条款符合。

附录 A
(资料性)

自身免疫病多组学数据库总体架构示意图

自身免疫病多组学数据库宜采用分层架构，如图 A.1 所示。各层之间通过受控接口连接，安全与质量控制贯穿全层，各层级组成见表 A.1。



图中为逻辑架构；实际部署可根据数据库规模和机构条件合并组件，但不得削弱职责分离、数据血缘和安全控制。

图 A.1 自身免疫病多组学数据库总体架构示意图

表 A.1 自身免疫病多组学数据库各层级组成

层级	主要组成
用户与治理层	治理委员会、数据责任人、伦理合规、临床科研用户、运维与审计人员
应用服务层	数据目录、队列构建、检索、分析、模型服务、共享审批、质量看板
数据治理层	主数据、元数据、数据字典、标准术语、质量规则、版本和血缘
数据资源层	临床、样本、组学原始数据、衍生数据、结果、模型、日志
技术平台层	数据库、对象存储、计算平台、工作流、接口、消息和监控
基础设施层	服务器、存储、网络、备份、灾备、机房或合规云资源
贯穿控制	身份权限、加密、审计、去标识化、完整性、备份恢复和持续监测

注：图 A.1 所示架构为逻辑示意。实际部署可根据数据库规模和机构条件合并组件，但不得削弱职责分离、数据血缘和安全控制。

附 录 B
(规范性)
数据库最小数据集及数据分类

数据库最小数据集及数据分类见表 B. 1。

表 B. 1 数据库最小数据集及数据分类表

数据域	最小内容	要求	敏感程度	管理属性
患者	患者研究标识、出生年份或年龄、性别、来源中心	必需	高敏感	严格受控
疾病	疾病名称与代码、诊断状态、诊断时间、病程、疾病活动度	必需	敏感	受控使用
治疗	药物、剂量、给药途径、开始/结束时间、疗效与不良事件	按用途必需	敏感	受控使用
样本	样本标识、类型、采集时间、处理、保存、质量、位置	必需	敏感	受控使用
检测	组学类型、平台、试剂、批次、运行编号、质控状态	必需	敏感	内部管理
原始数据	文件、格式、大小、校验值、生成时间和来源	必需	高敏感	严格受控
分析	流程、软件、参数、参考资源、任务和结果版本	必需	敏感	受控使用
随访结局	随访时间、疾病状态、复发、进展、器官损伤和结局	按用途必需	敏感	受控使用
授权伦理	同意范围、伦理编号、用途、共享和撤回状态	必需	高敏感	严格受控
质量审计	质量状态、问题、修复、访问、导出和共享日志	必需	视内容确定	内部管理
家系与遗传关系	家系标识、亲属关系、遗传关联和关系来源	按用途必需	高敏感	严格受控
注：敏感程度用于描述数据泄露、滥用或重识别可能造成的影响，可分为一般、敏感和高敏感；管理属性用于描述使用方式，可分为内部管理、受控使用和严格受控。机构应依据适用法律法规、数据规模、可识别性和应用风险形成正式分类分级目录，可提高敏感程度或管理强度，不应无依据降低。				

附 录 C
(规范性)
核心元数据和数据字典模板

核心元数据和数据字典模板见表 C. 1。

表 C. 1 核心元数据和数据字典模板

字段	说明	要求
数据元素标识	唯一编号	M
中文名称	业务可理解名称	M
英文名称	接口或数据库名称	M
定义	明确且无循环定义	M
数据类型	字符、数值、日期、布尔等	M
长度/精度	最大长度和小数位	C
格式	日期、编码或文件格式	C
值域/代码集	允许值及版本	C
单位	原始与标准单位	C
必填性	必填、条件必填或可选	M
空值含义	未采集、不适用等	M
来源	系统、表、设备或人工	M
转换规则	清洗、映射或换算	C
质量规则	规则编号和阈值	M
敏感级别	分类分级结果	M
责任人	数据责任人和管理员	M
生效版本	版本和生效日期	M
停用信息	停用日期和替代字段	C
业务定义版本	诊断标准、量表、结局定义或代码集版本	M
授权与用途	知情同意、伦理、共享、二次利用和撤回范围	M
时间语义	事件时间、采集时间、入库时间和时区	M
组学质量字段	按附录 L 记录的关键元数据和质量指标	C
注：M 表示必需，C 表示在适用时必需。		

附 录 D
(规范性)
患者、样本、检测批次和分析任务编码规则

D.1 编码原则

- 应按以下规定进行：
- a) 编码应在规定范围内唯一、稳定且不可重复使用；
 - b) 编码不应直接包含姓名、证件号、完整出生日期等直接标识信息；
 - c) 编码结构应有文档说明，结构变化应版本化；
 - d) 条码或二维码仅作为编码载体，不替代数据库校验；
 - e) 编码生成规则应规定字符集、长度、校验方法、生成主体和异常处置；
 - f) 编码一经分配不应重复使用，作废编码应保留状态和作废原因；
 - g) 跨中心共享时应保留本地编码和共享编码的映射关系，映射表应受控管理。

D.2 编码组成示例

患者、样本、检测批次和分析任务编码组成示例见表 D.1。

表 D.1 患者、样本、检测批次和分析任务编码组成示例（资料性）

对象	编码组成	示例
患者研究标识	中心代码-随机序列-校验位	AH01-X7K2P9-4
样本标识	患者研究标识-采样事件-样本类型-分装号	AH01-X7K2P9-4-V03-PB-02
检测批次	组学代码-日期-平台-序号	RNA-20260711-NOVA-003
分析任务	项目-数据版本-流程版本-序号	AIDB-DV2.1-PV1.4-018

注：表 D.1 仅为资料性示例，用于说明编码语义组成，不构成统一编码格式。机构可根据规模设定具体长度和字符规则，但应满足 D.1 的唯一、稳定、不可复用和可追溯要求。

附 录 E
(规范性)
数据质量维度、规则和错误代码

机构应至少建立本表所列质量维度和最低规则要求，并在规则库中规定适用数据域、阈值、严重性、责任人、处置时限和关闭条件。表 E.1 中错误代码仅为资料性示例，可由严重性、维度和序号构成：E 表示错误，W 表示警告，I 表示信息。

表 E.1 数据质量维度、规则和错误代码

维度	最低规则要求	错误代码示例
完整性	必填字段非空；关键时间点存在	E-COM-001
准确性	值与源系统或原始报告一致	E-ACC-001
一致性	性别与相关遗传/临床信息不矛盾；单位一致	E-CON-001
唯一性	患者、样本和任务编码无重复	E-UNI-001
及时性	数据在规定窗口内接入或更新	W-TIM-001
有效性	格式、值域、日期和文件校验值有效	E-VAL-001
可追溯性	结果关联源数据、流程和版本	E-TRA-001
可比性	中心、平台和批次差异已记录或校正	W-CMP-001
关联完整性	患者、样本、批次、分析任务和结果关联完整	E-REL-001
授权有效性	数据用途、伦理和知情同意状态在有效范围内	E-AUT-001
版本一致性	数据、流程、参考资源和模型版本完整且匹配	E-VER-001

附录 F
(资料性)
数据访问权限矩阵示例

数据访问权限矩阵示例见表 F. 1。

表 F. 1 数据访问权限矩阵示例（资料性）

角色	数据范围	允许操作	主要限制
临床医师	本人或授权患者的临床与报告数据	查询、补充	不得批量导出原始组学数据
实验人员	样本、批次、质控和检测数据	录入、查询	不得访问直接身份标识
数据管理员	元数据、字典、质量和版本	维护	不得审批自身数据申请
研究人员	批准项目的去标识化数据	查询、分析	仅在批准环境和期限内
系统管理员	系统和数据库配置	运维	默认不得浏览业务明文数据
安全审计员	访问和操作日志	只读审计	不得修改业务数据
外部协作方	协议范围内的最小数据集	受控访问	禁止再共享和本地长期保存

附 录 G
(资料性)
数据申请、审批、使用和退出流程

数据申请、审批、使用和退出流程见表 G. 1。

表 G. 1 数据申请、审批、使用和退出流程

步骤	环节	关键控制
1	提交申请	目的、项目、数据范围、人员、环境、期限和输出
2	完整性审查	材料、伦理、处理依据和利益冲突
3	数据与质量评估	数据可用性、质量、版本和适用性
4	安全与隐私评估	最小化、去标识化、环境和共享风险
5	审批与协议	权限、期限、成果披露、违约和销毁
6	数据准备	冻结版本、质量报告、最小数据集和标识
7	受控使用	日志、导出控制、进度和变更管理
8	成果审查	披露风险、引用和知识产权
9	退出关闭	撤权、返还/销毁、记录和问题结清

附录 H
(资料性)
备份恢复和灾难恢复检查表

备份恢复和灾难恢复检查表见表 H.1。

表 H.1 备份恢复和灾难恢复检查表

检查项目	结果
备份范围是否覆盖数据库、文件、配置、代码、字典和密钥恢复材料	☐ 是 ☐ 否
备份是否与生产环境隔离并实施加密	☐ 是 ☐ 否
是否定义 RTO、RPO 和最大可容忍中断时间	☐ 是 ☐ 否
是否定期验证文件校验值和备份完整性	☐ 是 ☐ 否
是否完成全量和增量恢复测试	☐ 是 ☐ 否
恢复后是否验证患者一样本—结果关联和权限	☐ 是 ☐ 否
是否开展灾难恢复演练并形成整改	☐ 是 ☐ 否
供应商退出或云故障是否有替代方案	☐ 是 ☐ 否
联系人、通信和应急文档是否为最新版本	☐ 是 ☐ 否
是否至少每年完成一次数据恢复验证	☐ 是 ☐ 否
临床应用型数据库是否至少每年完成一次灾难恢复演练	☐ 是 ☐ 否
不可变备份是否建立删除清单和恢复后再删除机制	☐ 是 ☐ 否

附 录 I
(资料性)
多中心数据映射与接入检查表

多中心数据映射与接入检查见表 I. 1。

表 I. 1 多中心数据映射与接入检查表

检查领域	检查内容
中心和来源系统登记	来源、责任人、时间范围和更新频率完整
患者匹配	本地标识、研究标识和跨中心映射已验证
疾病定义	诊断标准、活动度和结局定义一致或有映射
单位与参考范围	原始单位保留，转换规则和参考范围完整
样本流程	采集、处理、保存和质控差异已记录
组学平台	平台、批次、流程和版本可追溯
代码映射	本地代码与共同代码的映射经过复核
质量评价	完整性、一致性和分布差异已评价
安全合规	授权、伦理、传输、去标识化和共享协议有效
试运行	完成小批量接入、对账、回滚和性能测试

附 录 J
(资料性)
数据血缘和版本记录模板

数据血缘和版本记录模板见表 J. 1。

表 J. 1 数据血缘和版本记录模板

项目	记录内容
输出对象标识	结果文件、数据表或数据集的唯一标识
输出版本	版本号 and 发布日期
源数据	数据集、表、文件及版本
接入批次	接入任务或批次标识
转换规则	规则编号和版本
分析流程	流程名称、版本和容器/环境
软件参数	软件、版本和关键参数
参考资源	参考基因组、数据库、知识库和版本
运行信息	任务标识、开始/结束时间、计算环境
质量状态	通过、警告、失败及原因
责任人员	执行、复核和批准人员
变更记录	与前一版本的差异及替代关系

附 录 K
(规范性)
自身免疫病核心临床数据集

K.1 共性核心数据

数据库至少应按照表 K.1 建立共性核心临床数据。标记为“条件必需”的数据应在对应疾病、治疗或应用场景适用时采集。表 K.2 所列病种为代表性病种，不构成自身免疫病范围的封闭清单。

表 K.1 自身免疫病共性核心临床数据

数据域	数据元素	要求	说明
患者	患者研究标识	必需	不得包含直接身份标识
	出生年份或年龄	必需	保留计算日期或年龄基准时间
	出生性别/生理性别	必需	记录代码系统和未知状态
	身高、体重、体质指数	条件必需	记录测量时间和单位
疾病	疾病名称与代码	必需	采用标准术语并保留本地映射
	诊断状态	必需	疑似、确诊、排除或历史诊断
	诊断标准及版本	必需	记录标准名称、版本和满足项目
	发病时间和确诊时间	必需	未知时记录缺失原因
	疾病亚型或表型	条件必需	记录定义和版本
	疾病活动度量表	条件必需	保存量表名称、版本、项目和总分
	缓解/低疾病活动状态	条件必需	记录判定标准、时间和结果
	复发或疾病加重事件	条件必需	记录定义、起止时间和严重程度
器官受累	器官或系统受累	必需	记录起始时间、活动性和严重程度
	累积器官损伤	条件必需	记录量表、项目和总分
免疫学	抗核抗体	条件必需	方法、滴度/数值、核型、界值
	疾病相关自身抗体谱	条件必需	原始结果、方法、单位和界值
	补体 C3/C4 等	条件必需	原始值、单位、参考范围和时间
	免疫球蛋白及亚类	条件必需	原始值、单位、参考范围和时间
实验室	血细胞计数	条件必需	原始值、单位、参考范围和时间
	炎症指标	条件必需	如 ESR、CRP，保留方法和单位
	肝肾功能与尿检	条件必需	记录原始项目和参考范围
影像/功能	影像学检查	条件必需	检查类型、部位、日期、结论和原始文件标识
	肺功能、心功能或肌力等	条件必需	记录方法、设备、项目和单位
治疗	糖皮质激素暴露	条件必需	药物、剂量、途径、起止和累计剂量
	传统改善病情药/免疫抑制剂	条件必需	药物、剂量、途径、起止和停药原因
	生物制剂/靶向合成药物	条件必需	药物、批次或方案、起止和疗效
	治疗反应	条件必需	评价工具、时间点、项目和判定
	不良事件	条件必需	类型、严重程度、相关性、处置和转归
特殊事件	严重感染	条件必需	病原体、部位、住院、治疗和结局
	恶性肿瘤	条件必需	类型、诊断时间和治疗
	妊娠与生殖事件	条件必需	孕周、结局和药物暴露

数据域	数据元素	要求	说明
	疫苗接种	条件必需	疫苗、日期和不良反应
合并因素	吸烟、饮酒和职业暴露	条件必需	记录状态、量和时间
样本关联	采样时疾病状态	必需	活动度、治疗、感染和采样时间点
	采样时用药状态	必需	近期治疗和末次给药时间
随访	随访时间和方式	必需	门诊、住院、电话或其他
	疾病状态与结局	必需	稳定、缓解、复发、进展、死亡等
授权伦理	知情同意和伦理状态	必需	用途、共享、二次利用和撤回状态

K. 2 病种扩展数据

不同病种应在共性核心数据基础上增加病种特异性量表和结局。最低扩展内容见表 K. 2。

表 K. 2 常见自身免疫病病种扩展数据

病种	最低扩展内容
类风湿关节炎	关节计数、患者/医师总体评价、CRP/ESR、DAS28/CDAI/SDAI 构成项和总分、HAQ、RF、抗 CCP 抗体、影像学侵蚀及治疗目标状态
系统性红斑狼疮	SLEDAI-2K 构成项和总分、BILAG 系统评分、医师总体评价、SLICC/ACR 损伤指数、抗 dsDNA 抗体、补体、器官受累、缓解和低疾病活动状态
干燥综合征	ESSDAI 各领域和总分、ESSPRI 各项目 and 总分、腺体功能、抗 SSA/SSB 抗体、腺体影像或病理及系统受累
系统性硬化症	改良 Rodnan 皮肤评分、肺功能、HRCT 间质性肺病、肺动脉高压筛查、肾危象、指端血管病变和疾病相关自身抗体
炎性肌病	MMT-8、肌酶、肌炎特异/相关抗体、肌肉影像或病理、皮肤、肺和吞咽受累及治疗反应核心指标
抗磷脂综合征	血栓事件类型、部位和时间，妊娠并发症，狼疮抗凝物、抗心磷脂抗体和抗 $\beta 2$ 糖蛋白 I 抗体及其持续阳性状态，分类标准版本、抗凝治疗、复发和出血结局
系统性血管炎	血管炎类型和分类标准、器官受累、BVAS 或相应活动度工具、ANCA 及其他自身抗体、影像/病理、累积损伤、复发、感染和免疫抑制治疗
1 型糖尿病	疾病分期、胰岛自身抗体、空腹/餐后血糖、HbA1c、C 肽、胰岛素方案、连续葡萄糖监测、严重低血糖/酮症酸中毒及微血管和大血管并发症
自身免疫性甲状腺病	疾病类型、TSH、FT4/FT3、TPOAb、TgAb、TRAb、甲状腺影像、甲状腺眼病、药物/放射性碘/手术治疗及甲状腺功能结局
多发性硬化	疾病表型、复发时间和严重程度、EDSS、MRI 病灶部位和负荷、脑脊液寡克隆带、疾病修正治疗、残疾进展和认知/患者报告结局
重症肌无力	MGFA 临床分型、QMG、MG-ADL、AChR/MuSK/LRP4 抗体、胸腺影像或病理、肌无力危象、呼吸/吞咽受累、免疫治疗和治疗反应
炎症性肠病	克罗恩病/溃疡性结肠炎类型、病变部位和行为分型、疾病活动度、内镜和病理、粪钙卫蛋白、影像、并发症、手术、生物制剂/小分子治疗和黏膜愈合结局
乳糜泻	血清 tTG/EMA/DGP 抗体、总 IgA、HLA-DQ2/DQ8、十二指肠病理及 Marsh 分级、营养状态、无麸质饮食依从性、症状和并发症
自身免疫性肝病	疾病类型（如自身免疫性肝炎、原发性胆汁性胆管炎）、自身抗体、IgG/IgM、肝生化、影像和病理、纤维化/肝硬化分期、治疗反应、失代偿和肝移植结局
自身免疫性血液病	疾病类型（如免疫性血小板减少症、自身免疫性溶血性贫血）、血细胞计数、溶血指标、直接抗球蛋白试验、自身抗体、出血/血栓、输血、免疫治疗和复发结局
自身免疫性皮肤病	疾病类型（如白癜风、斑秃、天疱疮类疾病）、皮损部位和范围、严重度评分、皮肤病理/免疫荧光、相关自身抗体、治疗暴露、复发和患者报告结局

K. 3 其他病种扩展要求

未列入表 K. 2 的自身免疫病应在共性核心数据基础上建立病种扩展数据集，至少包括疾病诊断或分类标准及版本、疾病亚型、活动度或严重程度、主要器官功能、特征性自身抗体或生物标志物、影像或病理、治疗暴露、疗效和复发、并发症、患者报告结局以及病种特异性长期结局。数据元素的定义、单位、代码、评估频率和缺失状态应形成数据字典并保持版本可追溯。

附 录 L
(规范性)
不同组学类型最低元数据和质量字段

数据库应至少保存表 L.1 所列组学专项元数据和质量字段。质量阈值可由机构根据平台、用途和验证结果制定，但字段缺失、阈值版本和质量状态应可追溯。

表 L.1 各组学类型最低元数据和质量字段

组学类型	最低实验与技术元数据	最低质量字段
基因组/外显子组	样本类型、DNA 提取、建库方法、测序平台、读长、参考基因组、比对和变异流程版本	DNA 质量、测序深度、目标区域覆盖率、比对率、重复率、污染率、变异质控和失败原因
转录组	RNA 提取、RNA 类型、建库策略、链特异性、测序平台、读长、参考基因组和定量流程	RNA 完整性/质量、测序量、比对率、基因覆盖、rRNA 比例、重复率、样本相关性和批次状态
表观组	技术类型、前处理、建库、平台、参考基因组、峰或甲基化调用流程	转换效率或信号质量、测序深度、比对率、重复率、片段分布、峰数/覆盖和批次状态
单细胞组学	组织来源、消化方法、细胞/细胞核制备、活率、平台、目标细胞数、建库和分析流程	捕获细胞数、有效细胞数、双联体率、每细胞基因数/UMI、线粒体比例、环境 RNA、批次和细胞注释方法
空间组学	组织部位、切片编号、厚度、固定/保存、染色图像、平台、空间分辨率、坐标系统和配准方法	组织覆盖率、有效点位/细胞数、每点位基因数/UMI、图像质量、配准误差、背景信号和批次
蛋白质组	样本前处理、消化、分离、质谱平台、采集模式、内标、搜索库、鉴定和定量流程	仪器系统适用性、质控样本、肽/蛋白鉴定数、错误发现率、定量缺失率、重复性、批次漂移和异常批次
代谢组/脂质组	采样条件、前处理、内标、色谱和质谱平台、采集模式、化合物库、峰提取和注释流程	系统适用性、质控样本变异、保留时间/质量漂移、缺失率、批次校正、代谢物注释等级和空白污染
微生物组	样本部位、采集和保存、提取、扩增区域或宏基因组策略、平台、参考库和分析流程	阴阳性对照、测序深度、宿主污染、低生物量污染、读段质量、分类比例和批次状态
多组学整合分析	患者、样本和时间点配对关系，输入组学类型和数据版本，特征标识符映射，缺失模态处理，标准化与批次校正方法，整合算法、软件、参数和流程版本	跨组学配对完整性、输入样本数量、特征映射率、缺失模态比例、批次效应评价、整合稳定性/重复性、异常样本和失败原因

附 录 M
(规范性)
符合性评价项目、证实方法和判定规则

M.1 符合性评价

应至少覆盖表 M.1 所列关键项目。评价机构可增加项目，但不得减少与数据库用途直接相关的关键控制。评价方案应在实施前明确抽样对象、抽样数量、分层方式和判定阈值；患者、样本、检测批次、分析任务、结果、授权状态等关键关联应实施全量自动校验。现场人工抽样应覆盖不同中心、年度、病种和数据类型，每个关键数据域在每个被评价中心原则上不少于 30 条记录；总量不足 30 条时应全部核验。关键标识、关联、授权和版本字段完整率应达到 100%，其他适用必需字段完整率原则上不低于 95%，临床应用型数据库可根据风险规定更高阈值。

表 M.1 关键符合性评价项目

评价项目	主要证实材料	评价方法	合格判定
数据库用途与边界	用途说明、数据目录、禁止用途	文件审查、访谈	类型、用途、输出和限制明确且一致；临床应用型数据库具有附加验证和准入证据
治理与职责	组织架构、岗位职责、授权记录	文件审查、人员访谈	关键角色明确，无职责空缺或不可接受冲突；关键岗位具有有效授权
患者主索引	匹配规则、验证集、测试报告、合并拆分记录	规则复核、抽样测试	自动匹配规则经验证；抽样未发现错误自动合并；低于置信度阈值的记录 100% 进入人工复核
患者一样本一结果关联	样本链、批次和结果映射	全量自动校验、双向抽样追溯	关键关联全量校验通过；人工抽样无错误关联；任一错误关联按关键不符合项处理
核心临床数据	附录 K 数据字典和质量报告	字段抽样、语义核验	关键字段完整率 100%，其他适用必需字段完整率不低于 95%；量表版本、构成项和缺失原因可追溯
组学最低元数据	附录 L 字段和质控报告	文件与数据库对账	关键元数据和质量字段完整率 100%；非关键缺失已说明并限制用途
多组学整合	样本配对表、输入版本、整合流程和质量报告	配对核验、流程重放或结果复核	患者/样本/时间点配对关系明确，输入版本和算法可追溯，批次效应和整合稳定性已评价
数据入库	接入协议、校验和入库日志	批次抽查、重放测试	接收、校验、审核、版本和回滚证据完整；失败数据未进入正式可用状态
数据血缘	血缘图、流程、版本和参数	反向追溯测试	关键结果可定位源数据、流程、参数和参考资源，且可识别受影响的下游结果
数据质量	规则库、阈值、问题单和报告	规则复核、数据抽样	规则可执行、版本化；关键和重大问题已关闭；一般问题有经批准的整改期限
权限控制	权限矩阵、账号和审批记录	账号测试、日志核验	无越权访问；高权限账号至少每 6 个月复核一次，临床应用型数据库或高风险远程访问宜提高复核频次
去标识化与输出	处理方案、效果评价、导出记录	重识别风险测试	处理方式与用途匹配；稀有表型、家系和罕见变异输出采取聚合、抑制或人工审核
备份与恢复	备份记录、恢复报告、演练记录	实际恢复或见证演练	至少每年完成数据恢复验证；临床应用型数据库至少每年完成一次灾难恢复演练；数据、权限、关联和应用功能恢复完整
删除与销毁	审批、删除清单、销毁证明	记录核查、抽样验证	活动副本不可访问；不可变备份有删除清单和恢复后再删除机制；到期备份按计划销毁
模型训练数据	拆分清单、患者/家系标识、版本	脚本复核、抽样比对	患者级和适用时家系级数据泄漏为 0；训练、验证和测试数据版本完整
重大变更	变更单、影响评估、测试和回退	文件审查、现场核验	重大变更均经批准、测试和验证，具有可执行回退方案；变更后完成必要的专项评价
合规与伦理	处理依据、伦理、同意和人类遗传资源审查	文件审查、样本抽查	抽样数据用途与授权一致；撤回、限制和二次利用状态有效且能传递至衍生数据和共享对象

M.2 判定规则

存在关键不符合项，或存在未关闭的重大不符合项时，评价结论应为不通过。关键不符合项应立即采取控制措施；重大不符合项应在评价通过前关闭；一般不符合项应在经批准的期限内完成整改并

验证效果，原则上不超过 90 d。发生患者—样本—结果错误关联、未授权高敏感数据泄露、关键数据不可恢复或临床安全风险时，应直接判定为关键不符合项。

参考文献

- [1] 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
- [2] GB/T 20001.5—2017 标准编写规则 第 5 部分：规范标准
- [3] 中华人民共和国个人信息保护法（中华人民共和国主席令〔2021〕91号）
- [4] 中华人民共和国数据安全法（中华人民共和国主席令〔2021〕84号）
- [5] 中华人民共和国网络安全法（中华人民共和国主席令〔2016〕第 53 号）
- [6] 中华人民共和国生物安全法（中华人民共和国主席令〔2020〕第 56 号）
- [7] 网络数据安全管理条例（国务院令 第 790 号）
- [8] 中华人民共和国人类遗传资源管理条例（2024 年修订）（中华人民共和国国务院令〔2024〕777 号）
- [9] 个人信息保护合规审计管理办法（国家互联网信息办公室令 第 18 号）
- [10] 医疗卫生机构网络安全管理办法（国卫办规划发〔2022〕10 号）
- [11] 国家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）
- [12] National Academies of Sciences, Engineering, and Medicine. Enhancing NIH Research on Autoimmune Disease. Washington, DC: The National Academies Press, 2022. DOI: 10.17226/26554.
- [13] Mucke J, Bertias G, Aringer M, et al. EULAR recommendations for a core dataset to support clinical care and translational and observational research in systemic lupus erythematosus. *Ann Rheum Dis*, 2025, 84(10): 1621–1631. DOI: 10.1016/j.ard.2025.07.001.
- [14] Füllgrabe A, George N, Green M, et al. Guidelines for reporting single-cell RNA-seq experiments. *Nat Biotechnol*, 2020, 38(12): 1384–1386. DOI: 10.1038/s41587-020-00744-z.
- [15] Taylor C F, Paton N W, Lilley K S, et al. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol*, 2007, 25(8): 887–893. DOI: 10.1038/nbt1329.
- [16] Sumner L W, Amberg A, Barrett D, et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 2007, 3(3): 211–221. DOI: 10.1007/s11306-007-0082-2.