《增龄健康状态队列数据和样本采集处理指南》

团体标准编制说明

中山大学

广东食标科技有限公司

2025 年 10月

目 录

一. 标准制定的目的和意义

2021 年中国 65 岁及以上老年人口约 2 亿人，占总人口比重达 14.2%，中国正式步入老龄社会，而且预计未来几十年内老龄化现象将迅速加剧。根据预测，到 2050 年，中国 65 岁以上的人口将达到 3.95 亿，其中 80 岁以上的高龄老人将达到 1.35 亿。随着人口老龄化的持续发展，老龄化所伴随的慢性非传染性疾病的疾病负担也不断增加，包括心血管疾病、代谢性疾病、肿瘤、神经退行性疾病、骨骼/肌肉/关节疾病、衰弱等，是居民死亡的主要原因，给公共医疗体系带来巨大的压力。

老年人的健康状况受到整个生命周期中暴露因素累积效应的影响。因此，研究增龄健康状态的动态变化特征与规律，解析影响健康状态的关键因素和作用，开展早期干预来降低老年疾病风险提高健康水平，是实现健康老龄化亟需解决的关键科学问题。

增龄健康状态队列是指对同一研究人群在较长时期内进行多次、系统性的随访和数据收集，以追踪研究对象在年龄增长过程中生理、心理、社会功能等广泛健康状态的动态演变规律及其影响因素。

当前，我国在大跨度年龄健康研究领域仍存在空白，对增龄人群健康状态的动态演变过程的探索尚显不足。因此，建立和维护一个具有全国代表性、大规模、高质量的增龄健康状态队列，对于推动生物医学研究创新、实现健康增龄、为政府公共卫生决策制定提供科学依据具有重要的战略意义。它不仅能为自然科学领域发展做出实质性贡献，更有助于在全球生物医学研究中展现中国智慧和中国方案。

然而，大型人群队列研究具有随访周期长、任务重、覆盖区域广的特点，导致其数据来源广泛、内容复杂。规范、准确的数据是保障队列研究高质量和可持续性的关键前提。

本标准旨在解决上述挑战，针对增龄健康状态队列研究中的关键数据管理环节建立规范，具体包括：

1. 生物样本的采集、暂存、检测与保存；

2. 问卷数据的采集、处理与储存；

3. 多类型来源数据的整合与治理。

本标准将为不同队列研究的多源异构数据制定数据融合治理的标准化规范，以保证不同地区和研究条件下实现方法的一致性，也为新建队列的组织实施和数据管理提供重要的参考依据。

二. 标准的任务来源

2025 年 6 月，中山大学和广东食标科技有限公司，向广东省食品药品审评认证技术协会提出了制定《增龄健康状态队列研究数据和样本采集处理指南》团体标准的项目申请，正式提交了《增龄健康状态队列研究数据和样本采集处理指南》团体标准立项申请表。

三. 标准的主要工作过程

1.成立标准起草组

2025 年 5 月，广东省食品药品审评认证技术协会下达了该项目的制定计划任务，并组织中山大学和广东食标科技有限公司作为起草单位，成立标准起草组，对标准草案内容和整体框架进行了研讨，确定了工作进度时间表，并进行了分工，初步确定了标准框架和结构。

本标准起草单位有中山大学和广东食标科技有限公司，由广东省食品药品审评认证技术协会归口。

2.形成标准建议草案

2025 年 6 月，起草小组对标准的框架和内容进行了探讨，确定了制定计划、制定原则、标准框架、标准基本内容，并依此起草了《增龄健康状态队列研究数据和样本采集处理指南》。

2025 年 6 月-2025 年 7 月，起草小组对《保增龄健康状态队列研究数据和样本采集处理指南》初稿进行了深入的讨论，并征求了广东省内相关保健食品生产工艺专家的意见，对标准初稿进行了修改。

3.标准立项评估

2025 年 7 月，在前期研究的基础上，起草组根据 GB/T 1.1-2020 的要求进一步对标准内容和框架进行修改完善。2025 年 7 月 8 日，协会组织评审专家小组，召开团体标准立项评审会议，评审小组成员对该项目进行论证并获得通过。

4.形成标准征求意见稿

2025 年 7 月-2025 年 9 月，在立项审评会议的基础上，标准起草组召开内部研讨会，对标准的内容和架构进行了讨论和进一步修改。

起草组组织生物样本管理、数据统计与管理、信息安全等领域专家开展多轮研讨，细化各技术模块的具体要求。例如，针对数据采集、处理、储存的整体原则、生物样本采集与保存，针对问卷数据，规范采集内容（覆盖躯体生理功能、心理精神状态、社会适应能力）与质量控制方法（逻辑审核、缺失值编码）。经过 3 轮内部讨论与修改，形成标准初稿。

起草组也邀请了多位国家相关领导和专家，对团标进行了充分的讨论，对团标的建立给予了充分的肯定和支持，并提出了宝贵的意见和建议。标准起草组在反馈的建议，对团标的内容进行修改和完善，形成标准征求意见稿。

5.标准征求意见

2025 年 10 月-2025 年 11 月，《增龄健康状态队列研究数据和样本采集处理指南》征求意见稿面向行业内和社会各界公开征求意见。征求意见期结束后，标准起草组对收集回来的意见进行汇总，逐条进行讨论，作出采纳、部分采纳或不采纳的处理意见，并对团标相应的内容进行有必要的修改，形成送审稿。

6.标准审查

四． 标准制定依据和原则

1. 标准制定依据

编制《增龄健康状态队列研究数据和样本采集处理指南》主要参考了如下文件：

GB/T 1.1-2020《标准化工作导则》

ISO/IEC 27001:2013 信息安全管理体系

国际生物和环境样本库协会（ISBER）最佳实践

国家人口健康科学数据中心数据治理指南

2. 标准的制定原则

本标准制定遵循以下原则：

（1）规范性

按 GB/T 1.1-2020《标准化工作导则 第 1 部分：标准的结构和编写规则》的要求进行制定。

（2）一致性

与现行有效的相关法律法规一致，对增龄健康生物样本和问卷数据采集、数据整合等作出相应的规定。

（3）适用性

可以适用于广泛的老年人群的生物样本采集，问卷数据的采集，对此做出作出系统性的阐述和规定。

（4）可操作性

充分考虑当前老年人群的身体状况，采用已有的方法、指导原则等对标准进行规范，具有可操作性和普遍性。

（5）实用性

标准结合我国社区调研和临床实践特点设计可操作性条款，易于理解和执行，详细阐述样本和数据收集以及整合的步骤、具体内容和判定标准，使用方可直接参照执行，进行验策划、实施和结果输出。

五．标准的整体结构、主要内容及解析

本标准主要包括 12 个部分：范围、规范性引用文件、术语和定义、数据、生物样本采集和管理方案设计、数据信息采集、数据整合准备工作、数据标准化与一致性检查、数据清洗和检查、数据整合、生物样本采集、数据与样本储存要求。

标准的主要内容及解析如下：

1. 范围

本文件规定了增龄健康状态队列研究中数据采集、生物样品采集、数据与样本处理的基本要求和方法。

本文件适用于增龄健康人群队列研究的数据和生物样本的采集与处理。

2. 规范性引用文件

本标准参考了一项国际标准，ISO/IEC 27001:2013 《信息技术 安全技术 信

息安全管理体系 要求》，引用 ISO/IEC 27001:2013 作为数据安全管理的依据，因该标准是国际公认的信息安全管理权威标准，可确保数据存储与访问控制的合规性。

3. 术语和定义

对标准条文中出现的术语“增龄健康状态”、“问卷信息采集”、“生物样本”、“数据脱敏”、“样本标识系统”、“元数据”，在本标准中的含义作出了定义。

其中 “增龄健康状态” 定义为 “随着年龄增长，个体在生理、心理和社会功能等方面的健康状况综合体现”，参考《中国健康老龄化发展蓝皮书》中对 “健康增龄” 的界定，符合国内研究共识。其他几个术语均参考了相关标准描述。

4. 数据、生物样本采集和管理方案设计

作为增龄健康状态队列研究的 “顶层设计环节”，本章是后续数据采集、样本管理、质量控制的基础框架。大型人群队列研究具有随访周期长（通常 5-10 年）、覆盖区域广（可能跨多省市）、数据类型复杂（问卷 + 生物样本 + 体格检查） 的特点，若缺乏统一的管理方案，易出现 “各研究中心操作不一、数据无法溯源、样本质量失控” 等问题，直接影响研究结果的科学性与可重复性。因此，在研究启动前明确数据管理方案，是保障队列研究高质量推进的 “前置条件”。

4.1 方案设计要求：明确 “需涵盖多源数据质量管控”，因增龄健康研究需同步采集问卷调查（如生活方式）、生物样本检测（如血常规）、体格检查（如血压）等数据，不同来源数据的质量风险点不同（如问卷易出现填写错误，生物样本易降解），需在方案中提前规划针对性管控措施，避免后续数据整合时发现质量问题无法追溯。

4.2 四大核心原则：

标准化原则：针对国内队列研究 “各中心操作标准不统一” 的痛点（如部分中心血液离心转速为 2500rpm，部分为 3000rpm），统一操作流程与质量标准，确保不同地区、不同时间采集的数据具有可比性，为跨中心研究数据融合奠定基础。

安全性原则：既覆盖 “采样对象安全”（如采集前知情同意、避免操作损

伤），也包含 “样本生物安全”（如防污染、异常样本隔离），符合《中华人民共和国生物安全法》对生物样本管理的基本要求，同时避免因样本污染导致的研究资源浪费。

可追溯原则：建立 “样本标识 + 记录系统”，因队列研究需长期随访，样本可能在数年后用于检测，完整的追溯体系可快速定位样本来源（如受试者信息、采集时间），便于后续验证检测结果或排查质量问题（如某批次样本溶血，可追溯至采集人员或离心参数）。

质量控制原则：强调 “全过程监控”，而非仅在数据整理阶段质控，因生物样本（如粪便、口咽拭子）的质量受采集、处理、储存等多环节影响，仅靠后期质控无法挽回前期操作失误导致的样本失效（如粪便未及时低温保存导致核酸降解），需通过 “事前预防 + 事中监控 + 事后核查” 实现全链条质量管控。

4.3 生物样品流程图示：采用流程图形式呈现 “采集 - 处理 - 储存” 全流程，因生物样本处理涉及多个关键节点（如血液离心时间、样本转移时限），图示化设计可直观展示各环节衔接关系，降低基层操作人员的理解成本，避免因流程记忆偏差导致的操作失误（如漏离心、延迟冷冻）。

5. 数据信息采集

问卷数据是增龄健康队列研究中 “反映受试者主观状态与社会属性” 的关键数据（如生活方式、心理状态、社会支持），与生物样本数据、体格检查数据共同构成 “生理 - 心理 - 社会” 三维健康评估体系。本章规范问卷采集的全流程，解决 “调查工具不统一、采集内容不全面、质量控制缺失” 等问题，确保问卷数据的真实性、完整性与可比性。

5.1 采集方式：

专业培训与身份核对：增龄人群可能存在记忆偏差、听力 / 视力下降等问题，经培训的调查员可通过引导式提问（如 “您平时每周运动几次？”）提高回答准确性，同时核对身份可避免 “替答”（如家属代填）导致的数据失真。

纸质 / 电子问卷 + 经验证工具：纸质问卷适合无网络或老年人不熟悉电子设备的场景，电子问卷可实时逻辑校验（如填写 “从未吸烟” 后，自动隐藏 “戒烟时长” 选项），两种形式兼顾不同场景需求；“经验证工具”（如标准化的抑郁量表、认知评估量表）可确保问卷的信度与效度，避免自行设计问卷导致的

“问题表述模糊、维度缺失” 等问题。

跳转逻辑与无法回答记录：跳转逻辑可减少无效填写（如 “无高血压” 者无需填写 “高血压用药情况”），提高采集效率；记录无法回答的原因（如 “记不清”“拒绝回答”），可区分 “真缺失” 与 “逻辑空值”，避免后续数据清洗时误判，同时为分析数据缺失原因（如某问题拒绝率高，可能是表述敏感）提供依据。

5.2 采集内容：

三维度框架（躯体 + 心理 + 社会）：参考世界卫生组织（WHO）对 “健康” 的定义（生理、心理、社会适应的良好状态），覆盖增龄健康研究的核心维度，避免内容片面（如仅关注生理指标，忽略心理状态对健康的影响）。

躯体生理功能细化：包含人口学信息（如年龄、性别，用于分层分析）、疾病史（个人 + 家族，用于研究遗传因素影响）、健康行为（吸烟、饮酒，用于研究生活方式对健康的影响）、自评健康（受试者主观感受，补充客观指标的不足）、女性生育史（针对女性受试者的特殊健康影响因素，如绝经年龄与骨质疏松的关联），各模块均为增龄健康研究的经典变量，可支撑多维度分析需求。

心理精神状态：包含心理状况（如抑郁、焦虑）与认知状况（如记忆力、执行力），因增龄人群是心理问题（如老年抑郁）、认知障碍（如轻度认知功能损害）的高发群体，相关数据可用于研究认知衰退、心理问题的影响因素与演变规律。

社会适应能力：包含社会经济状况（如收入、教育水平，研究社会经济因素对健康的影响）、文化背景（如饮食习惯、地域文化，研究环境因素影响）、社会支持（如家庭支持、社区支持，研究社会互动对健康的保护作用），符合 “社会决定健康” 的研究理念，可丰富研究视角。

体格检查指标：选择身高、体重（计算 BMI）、腰围（评估中心性肥胖）、握力（反映肌肉功能）、血压、心率等基础指标，这些指标易采集、成本低，且与增龄相关疾病（如高血压、糖尿病、肌少症）密切相关，可作为健康状态的 “基础评估指标”，同时为后续衍生变量（如 BMI 分级）计算提供原始数据。

内容一致性要求：针对队列研究 “多轮随访” 的特点，要求不同中心、不同随访轮次的内容一致，避免因问卷内容变更导致的 “数据不可比”（如某轮

随访新增 “饮茶情况”，则无法分析基线至该轮的饮茶行为变化）。

5.3 质量控制：

现场初步检查：调查员在采集现场即时核对，可及时纠正明显错误（如年龄填写 “150 岁”）、补充遗漏（如漏填身高），避免 “样本回收后才发现问题，无法联系受试者补填” 的情况，提高数据完整性。

逻辑审核与一致性校验：数据整理阶段的 “二次质控”，通过逻辑规则（如 “出生日期晚于调查日期”“收缩压 < 舒张压”）排查矛盾数据，通过一致性校验（如同一受试者两次随访的性别不一致）发现录入错误，确保数据逻辑合理。

空值与缺失值编码：明确 “逻辑跳转空值保留空白，非跳题缺失值记为‘999’”，可避免后续分析时将两种空值混淆（如误将 “逻辑跳转空白” 视为 “缺失值”），同时统一缺失值编码便于统计软件识别与处理（如 SPSS 中可通过 “999” 快速筛选缺失值）。

6. 数据整合准备工作

数据整合是将 “分散的数据” 转化为 “可分析的统一数据集” 的关键环节，而 “备份原始数据” 与 “文档管理” 是数据整合的 “基础保障”，可防止数据丢失、确保整合过程可追溯。本章设置旨在解决 “原始数据未备份导致丢失”“处理过程无记录导致无法复现” 等问题，为后续数据标准化、清洗、整合提供安全与透明的操作环境。

6.1 备份原始数据：

多阶段备份：要求 “原始数据 + 每个处理阶段数据” 均备份，因数据整合涉及多次操作（如格式转换、编码映射），某一步骤失误（如误删除变量）可能导致数据损坏，多阶段备份可回滚至前一正确版本，避免 “一步错、全功弃”。

安全性与可追溯性：原始数据是研究的 “源头数据”，一旦丢失无法再生，备份可保障数据安全。

6.2 文档管理：

详细记录处理过程：要求记录 “处理依据、方法、结果”，因数据整合可能由多人协作完成（如 A 负责格式统一，B 负责逻辑校验），完整文档可确保不同人员理解操作意图（如 “为何将‘年’作为年龄单位”），避免因人员交接导致的操作偏差；同时，文档记录便于外部核查（如项目审计、论文评审），

证明数据处理的科学性与合规性。

透明性与可重复性：科研数据的核心要求是 “可复现”，即其他研究人员可根据文档记录重复数据整合过程，得到相同结果；文档管理是实现透明性与可重复性的关键，避免 “暗箱操作”（如未记录的异常值删除）导致研究结论不可信。

7. 数据标准化与一致性检查

增龄健康队列研究的数据来源多样（如不同中心的问卷、不同检测机构的生物样本报告），易出现 “数据格式不统一、编码不一致、逻辑矛盾” 等问题，导致数据无法整合分析。本章通过 “格式统一、编码映射、逻辑检查”，将异构数据转化为 “标准统一的数据集”，解决 “数据碎片化” 问题，为后续数据清洗、整合奠定基础。

7.1 数据类型与格式统一：

数据类型检查：确保 “数值型（如年龄）、字符型（如性别）、日期型（如采集日期）” 符合预期，因数据类型错误会直接影响分析（如将 “年龄” 误设为字符型，无法进行均值计算），提前检查可避免后续统计分析时的技术故障。

时间格式统一（hh:mm）：统一时间格式可避免因格式差异导致的排序错误（如 “13:00” 与 “1:00 PM” 无法正确排序），同时明确 “小时≤24，分 / 秒≤60” 的校验规则，排查无效时间（如 “25:61”）；日期合理性检查（如确诊时间≤当前年龄）可发现逻辑错误，确保时间数据真实可信。

单位标准化：采用国际单位制（如身高用 “m”，体重用 “kg”）与统一单位（如年龄用 “年”），避免因单位混乱导致的分析偏差（如部分中心身高用 “cm”，部分用 “m”，直接计算会导致结果相差 100 倍），同时符合国际科研数据交流的通用规范，便于研究成果国际化传播。

7.2 编码与映射：

统一分类数据编码：针对分类变量（如 “是否吸烟”“是否高血压”）制定 “1 = 是，2 = 否，3 = 不清楚” 的统一编码，避免不同中心自行编码导致的混乱（如 A 中心 “1 = 否，2 = 是”，B 中心 “1 = 是，2 = 否”），确保数据整合时分类变量含义一致，可直接用于合并分析（如跨中心统计 “吸烟率”）。

7.3 逻辑性检查：

时间顺序检查：如 “出生日期早于调查日期”“疾病确诊日期晚于出生日期”，这些是最基础的时间逻辑，违反则数据必然失真（如 “调查日期为 2025 年，出生日期为 2030 年”），需优先排查。

变量间逻辑关系检查：如 “从未吸烟” 者无 “戒烟时长”，“无子女” 者无 “子女赡养情况”，这类逻辑矛盾数据若不处理，会导致分析结果偏差（如将 “从未吸烟” 者的 “戒烟时长” 空值视为缺失，误算戒烟时长均值），通过逻辑检查可筛选并修正此类问题，确保数据内在一致性。

8. 数据清洗和检查

数据清洗是 “去除数据噪声、修正错误、补充缺失” 的关键环节，经过标准化与一致性检查的数据仍可能存在 “空值、异常值、重复值” 等问题，若直接用于分析，会导致研究结论偏差（如异常值拉高均值）。本章通过系统化的清洗措施，提升数据质量，确保数据满足后续统计分析（如回归分析、生存分析）的要求。

8.1 数据空值处理：

区分空值类型：明确 “缺失值（如受试者拒绝回答）” 与 “逻辑跳转空值（如无高血压者无需填写用药情况）”，避免将逻辑空值误判为缺失值，导致过度填补（如为 “无高血压者” 填补 “高血压用药” 数据），影响数据真实性；统一缺失值编码为 “999”，便于统计软件识别，同时与其他数值（如正常年龄、血压值）区分，避免混淆。

8.2 异常值检测和处理：

异常值识别：

统计方法（3 倍标准差法）：针对数值型数据（如年龄、身高、BMI），通过“均值 ±3 倍标准差”筛选超出合理范围的数据（如年龄 150 岁、身高 3m），该方法是统计学中识别异常值的经典手段，客观且可复现。

生理常识判断：结合增龄健康研究的受试者群体（通常为中老年人），设定合理范围（如年龄 30-100 岁，BMI 18.5-30），避免因统计方法的局限性（如数据分布非正态）导致的误判；选择题无效选项（如选项 1-3，填报 4）、单选题多答，均属于明显的录入或填写错误，需优先识别。

异常值处理：要求 “核查原因后决定保留 / 删除 / 替换”，而非直接删

除异常值，因部分异常值可能是真实数据（如极少数受试者 BMI>30，属于肥胖人群），盲目删除会导致样本损失；若确认是录入错误（如身高 170cm 误录为 1700cm），可修正后保留；若无法核查，标记为特殊值（如 “888”），便于后续分析时选择是否纳入，兼顾数据质量与样本量。

8.3 重复值检测：

重复行处理：重复行（如同一问卷录入两次）会导致样本量虚增，影响分析结果（如重复计算同一受试者的健康指标，拉高某指标的检出率），通过查找重复行并删除 / 合并，确保每一行数据对应一个唯一的观测（如一次随访的一个受试者）。

唯一标识符验证：为每个受试者分配唯一研究编号，避免因 “同一受试者多次随访编号不一致” 导致的重复记录（如基线编号 001，随访编号 002），确保 “一人一码”，便于纵向分析（如追踪同一受试者的健康状态变化），同时避免重复纳入同一受试者，保证样本代表性。

8.4 衍生变量计算与验证：

核查计算准确性：衍生变量（如 BMI = 体重 / 身高 ^{2}，血压分级 = 根据收缩压 / 舒张压划分）是后续分析的重要变量，计算错误会直接影响研究结论（如 BMI 计算错误导致肥胖分级偏差）；通过人工复核或软件自动校验（如用不同工具重复计算），确保衍生变量的准确性，避免 “一步错、步步错”。

9. 数据整合

数据整合是将 “清洗后的问卷数据、生物样本数据、体格检查数据” 合并为 “统一数据集” 的最终环节，也是实现 “多源数据联动分析” 的前提（如分析 “吸烟行为（问卷数据）- 炎症因子水平（生物样本数据）- 血压（体格检查数据）” 的关联）。本章明确整合范围与方法，解决 “数据分散存储、无法关联” 的问题，为研究人员提供完整的分析数据集。

9.1 数据整合范围：

问卷数据细化：再次明确 “躯体 + 心理 + 社会” 三维度的具体内容，与第 6 章采集内容呼应，确保整合时无数据遗漏（如不遗漏 “女性生育史”“社会支持” 等关键变量）；强调 “衍生数据”（如 BMI、血压分级），因衍生数据是后续分析的直接变量，需与原始数据一同整合。

生物样本数据：包含 “血常规、尿常规、血生化” 等基础检测指标与衍生数据（如血脂异常分级），这些指标是评估增龄人群生理健康的核心指标，与问卷数据整合后，可支撑 “生活方式 - 生理指标 - 健康结局” 的链式分析（如饮酒→肝功能异常→肝硬化风险）。

9.1.1 问卷数据整合

问卷调查数据可按照一般人口学信息、躯体生理功能、心理精神状态、社会适应能力进行整合。

躯体生理功能包括个人疾病史、家族病史、健康相关行为、自评健康状况、女性生育史；心理精神状态包括心理 状况、认知状况；社会适应能力包括社会与文化背景、社会经济状况、社会支持。

9.1.2 体格检查数据

体格检查数据主要包括身高、体重、腰围、臀围、握力、收缩压、舒张压、心率等基础数据，以及 由以上数据计算得到的衍生数据。

体格检查数据以躯体生理功能维度为主，数据包括基础指标和衍生指标。需要注意变量值间的逻辑核查，识别冲突值。对于异常数据应分析出现原因，通过通过核查原始问卷与数据集，重新计算衍生指标等方式修正错误。

9.1.3 实验室检查数据整合

实验室检查数据以躯体生理功能维度为主，包括基础指标、衍生指标。包括血常规、尿常规、肝功能、肾功能、血脂、尿酸、血糖、白蛋白等及各种衍生指标。

10. 数据质量控制

数据质量控制是贯穿增龄健康队列研究全流程的“保障机制”，前 11 章已涉及各环节的质量要求，本章通过“完整性、一致性、准确性”的系统校验，对数据质量进行“最终把关”，确保输出的数据集满足科研与应用需求（如论文发表、政策制定依据），同时为后续研究的质量改进提供依据。

10.1 完整性校验（MD5 校验）：

MD5 校验是一种哈希算法，可生成数据的唯一“数字指纹”，通过比对原始数据与备份数据的 MD5 值，可快速判断数据是否损坏或被篡改（如 MD5 值不一致，说明数据存在问题），该技术高效、可靠，是数据完整性校验的常用手段，

可确保数据在存储、传输过程中未发生改变。

10.2 逻辑一致性检查：

再次强调“时间顺序”（如出生日期早于调查日期）与“变量间关系”（如从未吸烟无戒烟时长），与第 8 章、第 9 章的逻辑检查呼应，属于“最终复核”，避免前期质控遗漏的逻辑矛盾数据进入分析环节，确保数据内在一致性。

10.3 空值、异常值、重复记录处理：

统一要求“非跳题缺失值编码为 999”“异常值用 3 倍标准差法识别”“重复记录确保唯一编号”，与第 9 章数据清洗要求一致，属于“质量验收标准”，确保各环节清洗措施执行到位，无遗漏问题。

10.4 异常值处理的统计与专业结合：

强调“统计方法（3 倍标准差）+ 专业知识”结合，避免仅依赖统计方法导致的误判（如某受试者 BMI=35，按 3 倍标准差可能被判定为异常，但结合临床知识，该值属于重度肥胖，是真实数据），确保异常值处理的科学性，平衡数据质量与真实性。

10.5 唯一研究编号检查：

再次核查“一人一码”，避免因编号错误导致的重复纳入或遗漏（如同一受试者有两个编号，导致被视为两人），确保样本代表性与分析准确性，尤其是纵向随访研究中，编号唯一性是追踪个体健康变化的基础。

10.6 过程记录与质量报告：

完整记录“清洗、修改、质控”过程，可追溯质量问题的原因（如某批次缺失值率高，记录显示是因调查员培训不足），为后续改进提供依据；定期生成质量报告（如缺失值率、异常值率、逻辑错误率），可直观展示数据质量水平，便于项目管理者把控研究进度，同时为研究成果的可信度提供证明（如论文中可引用质量报告说明数据质量）。

11. 生物样本采集与保存

生物样本是增龄健康队列研究的 “核心资源”，其质量直接决定后续实验室检测（如基因测序、生化指标分析）结果的准确性。本章针对队列研究中最常用的 4 类样本（血液、尿液、粪便、口咽拭子），细化技术要求，解决 “样本采集不规范、处理参数模糊、保存条件不明确” 等行业共性问题，确保样本质

量满足长期研究需求。

11.1 采集样本类型

分类明确样本类型：聚焦 4 类高价值样本，因这 4 类样本可覆盖多个研究维度（如血液用于生化指标检测，粪便用于肠道菌群分析，口咽拭子用于病原体检测），且采集难度较低、受试者接受度高，适合大规模人群队列研究。

细化采集参数：

血液样本 “空腹 10 小时以上”“EDTA 抗凝管 + 血清管各采 12mL”：空腹状态可避免食物对血糖、血脂等生化指标的干扰，两种管子分别用于不同检测需求（EDTA 管用于血常规，血清管用于生化指标），12mL 的采血量可满足多次检测需求（如部分样本需同时检测炎症因子、激素水平），避免重复采集对受试者造成负担。

尿液样本 “清晨第一次中段尿”“30 分钟内送检”：清晨中段尿污染物少、成分稳定，是临床常用的尿液检测样本类型；30 分钟内送检可避免尿液中细菌繁殖、化学成分分解（如尿蛋白变性），确保检测结果反映受试者真实生理状态。

粪便样本 “1g + 保存液”“剩余部分另存”：1g 样本量可满足肠道菌群测序等检测需求，保存液可维持样本活性，剩余部分另存为 “备份样本”，避免因单次检测失误（如样本污染）导致该受试者粪便样本完全丢失，保障随访研究的连续性。

口咽拭子 “双侧扁桃体各擦 6 次以上”“阴性对照”：扁桃体区域是病原体（如病毒、细菌）富集部位，6 次擦拭可确保采集到足量样本；设置阴性对照可排查操作过程中的污染（如拭子本身带菌），避免假阳性结果影响研究结论。

11.2 样本处理与保存：

强调 “尽快处理”：因生物样本（尤其是血液、口咽拭子）在室温下易发生降解（如血液中 RNA 降解、拭子中病原体失活），“尽快处理” 是保障样本质量的关键前提，后续条款中 “1 小时内转移至 - 80℃”“短暂 4℃保存” 均是 “尽快处理” 的具体落地措施。

血液离心参数（4℃、3000rpm、15 分钟）：该参数参考《临床生物化学检验血液标本的收集与处理》（WS/T 225—2024），可有效分离血浆 / 血清，同时避免离心转速过高导致红细胞破裂（引发溶血）或转速过低导致分离不彻底，

确保血清 / 血浆质量满足检测要求。

-80℃/ 液氮罐长期保存：-80℃冷冻箱和液氮罐可实现样本长期稳定保存（数年至数十年），符合增龄健康队列 “长期随访、多次检测” 的需求；相比 -20℃保存，-80℃可显著降低样本降解速率（如粪便中核酸 72 小时降解率从 30% 降至 5% 以下），保障样本在长期储存后仍具备检测价值。

11.3 样本标识与标签：

统一标识体系（条码 / 二维码）：相比传统手写标签，条码 / 二维码可避免人为书写错误（如编号写错、字迹模糊），且支持快速扫码入库，提高样本管理效率；管身 + 管盖双标注可防止单一标签脱落导致样本 “无主”（如管身标签磨损，管盖标签仍可识别）。

三级编码规则:“地区代码 + 队列类型 + 个体 ID + 样本类型 + 分装序号” 的编码结构，可实现样本 “唯一标识”，例如 “GD-A-001-B-01”（广东地区 + A 类队列 + 001 号受试者 + 血液样本 + 01 号分装），该编码可快速定位样本的地区、来源、类型，便于跨中心样本调拨与数据关联（如某受试者随访时的样本与基线样本通过个体 ID 关联）。

入库记录与数据关联：将样本记录与数据库关联，实现 “样本 - 数据” 一一对应，避免 “样本存在但无对应受试者信息” 或 “数据存在但样本丢失” 的情况，为后续 “样本检测结果 - 问卷数据” 整合分析提供基础。

11.4 样本采集注意事项：

知情同意与人员培训：知情同意是医学研究的伦理底线，符合《涉及人的生物医学研究伦理审查办法》；采集人员培训可确保操作规范性（如正确使用抗凝管、避免样本污染），减少因操作不熟练导致的样本质量问题。

污染防控与环境记录：样本污染是生物样本管理的主要风险（如粪便样本污染血液样本），记录环境条件（如室温、湿度）可排查污染原因（如湿度超标导致样本管受潮）；双人复核可降低单人操作的失误率（如漏记录采样时间）。

外观检查与质量评估：外观检查（如溶血、浑浊）可快速筛选明显异常样本，随机抽样检测（如核酸完整性）可评估批次样本质量，避免 “全员检测后才发现整批样本不合格” 的资源浪费；异常样本隔离与追溯机制，可防止问题样本扩散（如污染其他样本），同时定位问题环节（如某批次样本溶血，追溯至离心

操作），推动持续改进。

12. 数据与样本储存要求

数据与样本的长期安全储存，是增龄健康队列研究“可持续性”的关键——队列研究需长期随访（5-10 年），且样本与数据可能在数年后用于新的研究（如基因检测技术更新后，重新分析旧样本）。本章针对“储存规范、隐私保护、安全管理”提出要求，解决“储存不规范导致样本失效 / 数据丢失”“隐私泄露风险”等问题，保障研究资源的长期价值与合规性。

12.1 数据的储存要求

12.1.1 数据命名

统一命名规则：如“基线 / 随访_队列类型_疾病代号_组学类型_样本 ID”，该规则包含“时间（基线 / 随访）、来源（队列类型）、内容（疾病 / 组学）、标识（样本 ID）”关键信息，可快速识别数据 / 样本属性（如“随访 - A 队列 - 高血压 - 代谢组 - 001”，表示 A 队列高血压受试者随访时的代谢组数据），避免因命名混乱导致的查找困难（如“数据 1.xlsx”“样本 2.txt”无法识别内容）。

12.1.2 元数据管理：

分类变量编码统一：与第 8 章编码规则呼应，确保元数据中编码说明与实际数据一致（如元数据注明“1 = 是，2 = 否”，数据中编码也遵循此规则），避免后续使用时因编码理解偏差导致的分析错误。

标准化元数据文件：包含“采集时间、地点、方法、处理记录”，元数据是“数据的数据”，可解释数据来源与处理过程（如某批样本的检测方法是“高效液相色谱法”），便于后续研究人员理解数据背景（如判断检测方法是否适用于当前分析需求），同时符合科研数据“可追溯、可解释”的要求。

12.1.3 数据脱敏与隐私保护：

敏感信息脱敏：针对“姓名、身份证号、出生日期”等可识别个人身份的信息，要求“遮蔽 / 加密”，符合《中华人民共和国个人信息保护法》对敏感个人信息处理的要求，避免因数据泄露导致受试者隐私侵犯（如身份证号泄露引发诈骗）。

原始敏感数据单独加密：将原始敏感数据与脱敏后的数据分离存储，且仅授

权特定人员访问，可进一步降低泄露风险（如普通研究人员仅能访问脱敏后的“研究编号 + 健康数据”，无法获取身份信息）。

专人负责与复核记录：脱敏过程涉及数据修改，专人负责可明确责任，复核机制可避免脱敏失误（如误删非敏感信息），操作记录可追溯脱敏过程，确保合规性。

12.1.4 数据存储与备份：

安全可靠的存储环境：避免因存储环境差（如潮湿、高温）导致的硬件损坏（如硬盘故障），保障数据物理安全。

冷热三级存储体系：

热存储（SSD）：用于频繁访问的实时数据（如当前随访的原始数据），SSD 读写速度快，满足高频操作需求。

温存储（硬盘阵列）：用于近期需访问的数据（如近 1-2 年的处理后数据），硬盘阵列成本适中，兼顾速度与容量。

冷存储（磁带库）：用于长期归档的低频访问数据（如 5 年前的基线数据），磁带库容量大、成本低，适合长期保存。

该体系参考 ISO/IEC 27001:2013 信息安全管理标准，兼顾“访问效率、存储成本、数据安全”，符合大型队列研究“数据量大、访问频率差异大”的特点。

访问控制与双重认证：基于角色的访问控制（如研究人员仅能访问授权数据，管理员可访问全量数据）可避免越权访问；生物识别（如指纹）+ 密码的双重认证，可防止账号被盗导致的数据泄露，比单一密码更安全。

3-2-1 备份原则：“3 份副本、2 种介质、1 份异地”是国际通用的数据备份策略：3 份副本可避免单一副本损坏导致丢失，2 种介质（如硬盘 + 磁带）可避免同一介质故障（如硬盘批量损坏），1 份异地备份可防范本地灾难（如火灾、洪水）导致的全量数据丢失，全方位保障数据安全。

备份介质加密与密钥管理：备份介质（如磁带、移动硬盘）若丢失，未加密数据易被窃取，加密可防止数据泄露；专人管理密钥可避免密钥丢失导致备份数据无法解密，确保备份的可用性。

12.1.5 数据访问控制与安全管理：

最小权限原则：“按需分配权限”（如仅给统计分析师分配数据读取权限，不

给修改权限），可减少因权限过大导致的误操作（如误删数据）或恶意泄露风险。

保密协议与双重认证：保密协议明确数据管理人员的责任，可通过法律约束减少泄露行为；访问系统的双重认证，进一步强化身份验证，确保只有授权人员可进入系统。

日志审计与定期审查：记录用户操作（如“谁在何时访问了哪些数据、是否修改”），可追溯异常操作（如某用户频繁下载敏感数据）；定期审查日志可及时发现安全隐患（如未授权访问尝试），防范数据泄露。

针对样本储存和运输标准起草组也提出了相关条款建议，相关条款建议如下

12.2.1 存储条件

短期存储若使用 4℃环境储存，储存时间不宜不超过 24 小时，-80℃存储不宜超过 5 年。

长期存储可采用液氮气相长期保存。

12.2.2 存储管理

可采用 24 小时温度监控与报警系统，进行双人双锁管理制度，每月库存盘点与质量抽检。

12.2.3 样本运输

运输容器宜使用认证的低温运输箱，干冰量应保证至少 5 天维持温度。

运输要求需全程进行温度监控与记录，交接时检查样本状态并记录。

六．与国内相关标准的关系

本标准与现行法律、法规、规章及相关标准无冲突，具体协调性分析如下：

（1）与法律、法规的协调性：符合《中华人民共和国个人信息保护法》对个人敏感信息处理的要求（如数据脱敏、加密存储），符合《中华人民共和国生物安全法》对生物样本采集、保存、使用的生物安全管理规定，确保标准内容合法合规。

（2）与国家标准的协调性：与《生物样本库质量和能力通用要求》（GB/T 37864—2019）、《卫生信息数据元标准化规则》（WS/T 303—2009）、《数据质量评估指南》（GB/T 35273—2020）等国家标准保持一致，例如生物样本标识

要求与 GB/T 37864—2019 中的“唯一标识”原则一致，数据元规范与 WS/T 303—2009 中的编码规则一致。

（3）与行业标准的协调性：与《临床生物化学检验血液标本的收集与处理》（WS/T 225—2021）、《成人体格检查规范》（WS/T 427—2013）等行业标准衔接，确保生物样本采集、体格检查指标等技术要求与临床实践、行业惯例一致，避免标准间的技术冲突。

七．重大分歧意见的处理经过和依据

本标准在制定过程中未出现重大分歧意见。

八．专利及涉及的知识产权

本标准不涉及任何专利或知识产权。

九．标准意见汇总处理表