ICS 11.020
CCS C05

团 体 标 准

T/BRACDCHE 004-2025

跨队列研究数据采集质量控制要求

Requirements of Quality Control on Data Collection for Cross-cohort Study

2025 - 05 - 26 发布　　2025 - 05 - 26 实施

北京慢性病防治与健康教育研究会　　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。

本文件由北京大学第六医院提出。

本文件由北京慢性病防治与健康教育研究会归口。

本文件起草单位：北京大学第六医院、天津市安定医院、北京大学、中国医学科学院肿瘤医院、山东大学齐鲁医院、中国疾病预防控制中心、北京大学第一医院、中国电子技术标准化研究院。

本文件主要起草人：刘肇瑞、徐广明、尹慧芳、黄雨、罗雅楠、黄悦勤、魏文强、吕明、陈园生、张婷婷、丁若溪、邓咏妍、李明慧、孙可欣、李航、李瑞琪、杨孝荣、陈浩、白倩倩、王悦。

跨队列研究数据采集质量控制要求

1 范围

本文件规定了在开展跨队列研究前，原始队列数据的采集内容及方式和质量控制要求等。

本文件适用于跨队列研究原始队列数据采集工作的质量控制，包括但不限于社区人群队列、区域性人群队列、针对某一疾病种类或基于特殊暴露因素建立的人群队列。

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 39767-2021 人类生物样本管理规范

WS/T 306 卫生信息数据集元数据规范

WS 363 卫生信息数据元目录（所有部分）

WS 364 卫生信息数据元值域代码（所有部分）

WS 365 城乡居民健康档案基本数据集

WS/T 370 卫生信息基本数据集编制规范

WS 372.3-2012 疾病管理基本数据集

WS 375 疾病控制基本数据集（所有部分）

T/CPMA 004-2019 大型人群队列终点事件长期随访技术规范

3 术语和定义

以下术语和定义适用于本文件。

3.1

队列 cohort

根据某个或某些共同特征而组建的一组特定人群。

注：特征包括但不限于：暴露因素、疾病或健康状态、出生时间或年代、地域、干预措施等。

3.2

跨队列 cross-cohort

队列（3.1）间进行特征数据比较、融合和分析。

注：跨队列形式包括：a）横向跨队列：在不同元数据的队列间进行比较、融合和分析；b）纵向跨队列：在相同元数据的队列间进行比较、融合和分析。

3.3

调查对象 participants

队列（3.1）中的人。

注：可能是一般人群，也可能是患者或具有某些危险因素的人群。

3.4

病例报告表 case report form

按照研究方案要求设计的记录调查对象相关信息的纸质或者电子文件。

3.5

标准操作程序 standard operating procedures

为保证队列数据采集工作的一致性而制定的详细的数据采集程序。

3.6

随访终点 end point

研究对象在随访期间内出现了预期结局（如死亡、发病等）或者失访，即为随访终点。预期结局根据其研究目的进行设定。

[来源：T/CPMA 004-2019]

3.7

终点事件 endpoint events

终点事件是临床研究或试验中根据研究目的设定的关键事件或指标（如发病、死亡等），用来评估干预措施的效果。肿瘤登记随访的终点事件一般为死亡。

[来源：T/CPMA 004-2019]

3.8

被动随访 passive follow-up

被动随访由指定的机构利用当前运行的各类监测系统或常规工作中形成的资料或数据库,以间接方式定期或不定期采集到研究对象的终点事件的方法。

3.9

主动随访 active follow-up

主动随访是由负责随访的人员，主动收集患者的信息资料及（或）医院、医生与研究对象本人或其亲属保持常规的接触，了解他们当前的状态，调查研究对象的病情或生存状况，并归纳整理。

3.10

最后接触日期 date of last contact

为最后知晓患者生存状态的日期。

注：如果已知患者死亡，最后接触日期应为死亡日期。如果患者失访，最后接触日期应为最后一次知晓患者生存状态的日期。如果患者存活，最后接触日期应为随访当天日期。

3.11

最后接触状态 condition of last contact

为最后知晓的患者的生存状态，包括存活、死亡或失访。

3.12

死亡原因 cause of death

所有导致或促进死亡的疾病、病态情况或损伤以及造成任何这类损伤的事故或暴力情况。

注：该定义不包括症状、体征和临床死亡方式（如心力衰竭或呼吸衰竭）。

[来源：T/CPMA 004-2019]

3.13

根本死因 underlying cause of death

直接导致死亡的一系列病态事件中最早的那个疾病或损伤,或者是造成致命损伤的那个事故或暴力情况。

[来源：T/CPMA 004-2019]

3.14

迁移 migration

研究对象的户口迁出调查区域。

注：对于迁移的研究对象，由最新户籍地的登记处负责进行管理和随访工作。

[来源：T/CPMA 004-2019]

3.15

研究对象的失访 loss to follow-up

研究者无法对研究对象进行长期随访以获得终点事件信息的情况。

注：失访的原因主要包括拒访、搬迁、失联、查无此人等。

3.16

结局事件 outcome events

结局事件是在研究中发生的任何与患者健康相关的事件，它包括但不限于治疗效果。

注：结局事件可以是积极的（如疾病的缓解）或负面的（如不良反应、疾病进展等）。可以是疾病发生，健康情况改变，或死亡等。

3.17

结局事件发生日期 date of outcome event

在随访过程中观察到的结局事件出现的时间。

4 采集内容及方式

4.1 采集内容

采集内容为调查对象的社会人口学数据、疾病或健康数据、医疗和照护信息，文化背景、社会支持情况，遗传学信息以及与疾病和健康有关的生活方式和行为等信息。

4.2 采集方式

采集方式为自评问卷、他评问卷、体格检查、辅助检查（包括影像学检查以及生物样本检测等）。

注：宜为每个调查对象确定唯一标识符，例如身份证号、社会保障号、样本编号或病历号，以便数据清理、整合和分析。

数据应符合卫生健康相关数据集的标准WS/T 306、WS 363、WS 364（所有部分）、WS 365、WS/T 370、WS 372.3-2012、WS 375（所有部分）。

5 质量控制要求

5.1 数据采集前

5.1.1 操作程序

应明确数据采集流程，确定数据质量控制评估标准，并建立质量核查机制。质量控制标准中应包含量化指标，以确保数据质量达到既定要求。

5.1.2 病例报告表

应确定病例报告表的内容，开发完成数据采集或录入系统，并对数据录入格式及有效值范围进行设定，尽量使用电子化病例报告表，一方面可采集并行数据以进行质量核查，另一方面可减少录入错误以提高数据质量。需要根据研究目标确定队列数据的特有信息，内容包括随访终点、终点事件、被动随访、主动随访、最后接触日期、最后接触状态、死亡原因、根本死因、迁移、失访、结局事件和结局事件发生时间等。

5.1.3 人员培训

a) 培训对象

培训对象为所有参与数据采集和质量控制的人员。培训内容包括基本职责培训和专业技能培训，经考核合格后方可参与数据采集或质量控制工作。

b) 培训流程

数据采集和质量控制工作应由不同的人员开展。

所有数据采集和质量控制人员还应参加定期培训和阶段性考核并成绩合格，确保所有人员持续具备良好的数据采集能力，从而保证队列建设高质量地推进。同时，可根据研究要求，对参与人员的资质进行评估，内容包括但不限于其是否熟悉当地语言、生活方式和习惯，是否具有相关医学教育背景、职称和工作经历等。

5.1.4 现场检查

评估调查现场准备妥当，符合调查的标准，并对现场评估情况进行记录，现场评估按照以下内容进行：

a) 现场整洁，秩序良好，私密性强，确保调查不被干扰；
b) 项目标识和引导牌是否明确；
c) 是否设置现场负责人和协调员；
d) 采集人员是否佩戴工作胸卡或身穿统一服装；
e) 调查所用工具是否准备齐全且完好无损可使用；
f) 实验台面、耗材摆放是否干净且整齐；

g) 如使用电子化数据采集，还应检查采集设备及采集程序是否可正常使用。例如硬件设备的功能性检查、软件程序的兼容性和稳定性测试等，确保所有设备在采集过程中处于正常工作状态，并能准确记录数据。

5.2 数据采集中

5.2.1 核实调查对象情况

核实调查对象符合/纳入标准，不符合/排除标准，是否具有采集信息的条件，以及调查对象是否签署知情同意等。

5.2.2 采集信息

由数据采集人员按照标准操作程序的数据采集流程采集信息，注意保护调查对象的隐私，主观访谈时不诱导或代替调查对象回答问题，客观检查时应注意保存检查的原始结果。重点采集队列数据特有信息，内容包括但不限于随访终点、终点事件、被动随访、主动随访、最后接触日期、最后接触状态、死亡原因、根本死因、迁移、失访、结局事件和结局事件发生时间等。参与者的姓名、联系方式、死亡原因等敏感信息应加密存储，且仅限于研究团队成员访问。

5.2.3 及时检查数据质量并进行反馈和整改

由质量控制人员按照标准操作程序的数据质量控制评估标准以及质量核查方式，及时检查数据质量。设置在调查对象数据采集后特定时间内完成数据质量检查。数据检查应设定具体的量化标准，指标包括但不限于错误率、数据完整率。如发现存在质量问题，应在规定时间内反馈给数据采集人员，并监督其完成整改。对存在严重质量问题的数据，应重新采集或对该条数据进行废弃标记。对数据进行核查包括但不限于以下内容。

a) 数据唯一性：通过比较每名调查对象收集的可识别数据项（如姓名、身份证号、社会保障号）进行核查，可核查数据集内或数据集间不同研究对象的个体唯一性标识和有效记录是否重复。
b) 数据完整性：核查数据集的实际样本量或记录数与应获得数量是否相同；核查变量的完整性，除外重复变量，核实数据集中已有变量数与应获得变量数是否相同；核查变量值的完整性，如数据集中特定单元格信息是否完整，是否有缺失的数据，如有，应明确记录数据缺失的原因。
c) 数据合法性：对单一数据元进行核查，是否有非法编码，是否超出数据元属性。
d) 内部一致性：对多个数据元之间的逻辑进行核查，是否存在逻辑不合理性（如确诊时间晚于调查对象的年龄，性别不符等）。

5.2.4 核查调查的真实性

质量控制人员可通过多种方式判断调查的真实性。核查的方式包括但不限于以下内容：

a) 数据核查：对数据采集的时长和速度信息进行监控，如时长过短速度过快则有可能调查不真实；
b) 电话核查：致电调查对象，了解调查时的相关情况，并对关键人口学信息进行比对，如调查对象否认接受调查，或关键信息比对失败则有可能调查不真实；
c) 录音核查：在获得调查对象允许的前提下对调查过程进行录音，如录音率过低或录音中声音与调查对象年龄或性别不符则有可能调查不真实；
d) 实地核查：选取部分调查对象，再次进行调查或进行访谈以了解调查过程，如调查主要结果不符或调查对象否认参与调查则有可能调查不真实。

5.2.5 定期评估反馈并作调整

定期生成执行进度报告和质量控制报告，由项目负责人团队对现场采集的数据进行描述性统计和分析，发现现场执行中的问题（如执行率低、失访率高等），查找系统误差产生的原因，向数据采集人员和质量控制人员反馈，传达调整建议并督促落实。

5.3 数据采集后

5.3.1 数据备份和安全

在数据清理前，应先对原始数据进行备份，确保数据的完整性和可追溯性。数据的可追溯性应通过记录数据备份的版本号及操作日志来量化和监控。数据备份应包括多个版本，以防数据损坏或丢失。备份数据应存储在安全的环境中，并设置严格的访问权限。数据备份完成后再进行数据清理。数据备份应确保所有版本在多个物理位置进行存储，并且备份数据的访问应有严格的权限控制，包括但不限于设置访问记录、备份操作日志和定期验证备份的完整性。

5.3.2 数据清理

数据清理时不可在原始数据库中直接操作，应设定数据清理原则，通过编写数据清理命令进行数据清理，并将清理后的数据另存为新的数据库。数据清理命令中应有进行后期核查的注释性文字。每一步操作后保存一个具有明确命名的中间文件，以便回朔操作流程。

数据清理前应先进行数据完整性审核，确保所有收集的数据已经通过初步的质量检查，包括缺失数据、重复数据和异常值的初步筛查。

数据清理时应包括检查和处理以下内容：

a) 缺失数据处理：如某变量的缺失可能导致主要结局受影响，则应删除关键变量缺失的行或使用插补方法补全缺失数据；根据研究目的和具体情况设置缺失数据的删除条件；
b) 异常值检查和处理：利用统计方法设置异常值识别标准并决定是否修正或删除；
c) 重复值去重：识别和删除重复记录或合并重复信息；
d) 数据一致性检查：确保同一调查对象的不同信息之间没有逻辑冲突。

5.3.3 完善质量监控体系

质量控制体系应包括定期的审查和更新机制，对数据采集和处理的各个阶段进行质量评估和反馈。在监控中，建议明确具体的技术路线，并在每个阶段标注关键量化指标（包括但不限于数据录入错误率、异常值比率），以实现精准的质量控制。建立质量控制报告制度，对发现的问题应及时记录、反馈，并制定整改措施。质量控制报告应定期由项目管理团队审阅，以确保质量控制工作持续改进。

5.3.4 锁定数据库并撰写数据元专用属性表

数据清理后锁定数据库。撰写数据元专用属性表，即数据字典，内容包括但不限于数据元名称、定义、数据元值的数据类型、表示格式以及数据元允许值。

5.3.5 文件归档

及时将知情同意书、调查问卷、病例报告表、临床检验或检查报告单等原始资料归档保存，纸质版原始材料应由专人加密保管，档案保存时限不少于10年。在确保安全的前提下，可以实行电子归档。

5.3.6 生物样本的保存、运输和使用

如采集了生物学样本，生物样本的保存、运输和使用严格遵循GB/T 39767-2021的要求。

5.4 随访期间质量控制

5.4.1 终点事件的确定

确认每个队列研究所关注的终点事件，包括主要结局与次要结局。明确并统一结局变量的设置与编码，确保不同队列和研究者之间的一致性与可比性。所有结局事件应有明确的定义，并与临床指导原则保持一致，确保数据收集的标准化。

5.4.2 随访方式

确认每个队列研究开展随访的具体途径，包括依托死因监测、专病监测等系统开展的被动随访，或主动随访，或被动、主动随访均开展。每种随访方式应明确其适用条件、频率和操作流程，随访的频率应根据研究设计的要求进行标准化。

5.4.3 完整性审核

审核不同队列随访信息的完整性，完整率宜高于95%。

5.4.4 失访率的要求

队列研究的失访率不宜高于15%。当失访率高于15%时，应详细调查失访的原因。研究团队应对失访事件进行详细记录，并由专家组评估失访率较高对随访数据质量和数据融合质量的影响。失访率过高时，应考虑重新调整调查策略。

参 考 文 献

[1] 龚巍巍, 俞敏, 郭彧, 等. 大型人群队列终点事件长期随访技术规范团体标准解读[J]. 中华流行病学杂志, 2019, 40(7): 756-758.