ICS 11.020
CCS C05

团 体 标 准

T/BRACDCHE 005-2025

跨队列研究质量控制数据元定义及编码要求

Definitions of Quality Control Element and Coding Verification 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 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　质量控制数据元模型 2
4.1　组成 2
4.2　版本控制 2
5　数据元定义及编码 2
5.1　队列数据基本信息数据元 2
5.2　队列数据研究对象信息数据元 3
5.3　观测指标和变量信息数据元 4
5.4　数据采集信息数据元 4
5.5　随访信息数据元 5
参考文献 6

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、北京大学、中国医学科学院肿瘤医院、山东大学齐鲁医院、天津市安定医院、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、黄雨、魏文强、孙可欣、岳伟华、张婷婷、黄悦勤、吕明、张媛、徐广明、陈园生、罗雅楠、丁若溪、邓咏妍、李航、尹慧芳、侯筱菲、李瑞琪、殷晓霖、白倩倩、李泊萱、王悦。

跨队列研究质量控制数据元定义及编码要求

1 范围

本文件规定了在开展跨队列研究时，原始队列数据集质量控制数据元定义及编码要求等内容。

本文件适用于开展跨队列研究的原始队列数据集的质量，包括社区人群队列、区域性人群队列、针对某一疾病种类或基于军队、学校等特殊机构建立的人群队列。

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 11643 公民身份号码
GB/T 2260 中华人民共和国行政区划代码
GB/T 2261.1 个人基本信息分类与代码第1部分：人的性别代码
GB/T 2261.2 个人基本信息分类与代码第2部分：婚姻状况代码
GB/T 2261.4 个人基本信息分类与代码第4部分：从业状况(个人身份)代码
GB/T 3304 中国各民族名称的罗马字母拼写法和代码
GB/T 4658 学历代码
GB/T 14396 疾病分类与代码
GB/T 18391.1-2009 信息技术 元数据注册系统(MDR)
GB 18030-2022 信息技术 中文编码字符集
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 372-2012 疾病管理基本数据集
WS 375 疾病控制基本数据集（所有部分）

3 术语和定义

下列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：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

数据元 data element

由一组属性规定其说明、标识、表示和允许值的数据单元。

[来源：GB/T 18391.1-2009,3.3.8]

3.4

字符　character

一个对象或实体的特性，供组织、控制或表示数据用的元素集合中的一个元素。

[来源：GB 18030-2022]

4　质量控制数据元模型

4.1　组成

队列数据集的质量控制数据元模型由队列基本信息、队列数据研究对象信息、观测指标和变量信息、数据采集信息和随访信息五组数据元组成。

4.2　版本控制

详细记录每次版本更新的内容，包括更新的日期、修改的具体内容、修改原因及相关责任人。建立有效的文档管理机制，确保所有版本能够被及时存档和检索，方便后续的审查和使用，确保数据的长期管理和维护。

5　数据元定义及编码

5.1　队列数据基本信息数据元

5.1.1　队列名称

队列数据的名称，文本数据，可为中文或英文，长度不限，建议不超过50字符，名称中应说明队列数据的采集地点和研究对象的疾病或健康属性。

5.1.2　队列标识符

队列数据的唯一标识，文本数据，由英文、数字或特殊字符组成，长度不限。

5.1.3　研究目的

描述基于研究问题、研究假设开展研究的意义、目的、研究问题、研究假设等，文本数据，可为中文或英文，长度不限。

5.1.4　研究背景

描述研究的背景信息、研究动机、相关领域情况介绍等，文本数据，可为中文或英文，长度不限。

5.1.5　研究范围

描述研究的范围、涵盖的主题、时间跨度等，文本数据，可为中文或英文，长度不限。

5.1.6　研究样本量估计

描述研究样本量计算过程，文本数据，可为中文或英文，长度不限。

5.1.7　研究样本量

说明研究样本量，数值信息，为不超过32个字符的整数。

5.1.8　数据收集时间

包括数据收集开始和结束时间，即收集开始和结束的年、月、日（YYYY-MM-DD），日期变量，为10个字符。

5.1.9　研究团队

描述参与研究的研究人员、研究机构等，文本数据，可为中文或英文，长度不限。

5.1.10　数据来源

描述队列数据的来源，包括资金来源、数据归属机构等，文本数据，可为中文或英文，长度不限。

5.1.11 数据存储和管理

描述数据存储、管理、保护的方式、策略等，文本数据，可为中文或英文，长度不限。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）。

5.1.12 隐私和伦理审查

描述与研究中隐私保护和伦理规范相关的信息，文本数据，可为中文或英文，长度不限。

5.2 队列数据研究对象信息数据元

5.2.1 身份识别信息

可识别研究对象身份的信息，包括身份证号、医疗保险号、病历号等，文本数据，按照WS364.3 CV02.01.101 身份证件类别代码表、 GB 11643规定执行。

注： 输入的个人信息会被加密后存储在政府机构的数据库中，只有授权的工作人员才能查看和使用这些信息。

5.2.2 研究对象编号

队列数据内部使用的研究对象唯一号码，应与身份识别信息一一对应，但不包括身份识别信息，数值数据，长度不限。

5.2.3 姓名

包括研究对象本人当前在公安户籍管理部门留档、正常使用的姓氏和名字，文本数据，最大20个字符。

5.2.4 出生日期

研究对象出生当日的公元纪年日期的完整描述，出生当日的年、月、日（YYYY-MM-DD），数值数据，为10个字符。

5.2.5 年龄

研究对象参与队列数据采集时的年龄，数值数据，为不超过3个字符的整数。

5.2.6 性别

研究对象生理性别在特定编码体系中的代码，数值数据，取值范围为0,1,2,9，应符合GB/T 2261.1的规定。

注： 根据GB/T 2261.1的规定，研究对象的生理性别代码通常定义如下：0：未知性别；1：男性；2：女性；9：其他（或未说明的性别）。

5.2.7 教育水平

研究对象参与队列数据采集时接受教育的最高程度类别在特定编码体系中的代码，数值数据，最大2个字符，应符合GB/T 4658相关标准。

5.2.8 民族

研究对象所属民族在特定编码体系中的代码，数值数据，最大2个字符，应符合GB/T 3304相关标准。

5.2.9 婚姻状况

研究对象参与队列数据采集时的婚姻状况在特定编码体系中的代码，数值数据，取值范围为01，02,03,04，05，最大2个字符，应符合GB/T 2261.2的规定。

注： 根据GB/T 2261.1的规定，研究对象的婚姻状况代码通常定义如下：01未婚；02已婚；03离婚；04丧偶；05其他。

5.2.10 职业类别

研究对象参与队列数据采集时的从事最长时间的职业类别在特定编码体系中的代码，数值数据，最大2个字符，应符合GB/T 2261.4的规定。

5.2.11 出生地

研究对象出生时所在地区对应的6位行政区划代码，数值数据，最大6个字符，应符合GB/T 2260的规定。

5.2.12 常住地

研究对象参与队列数据采集时常住的地理位置，包括国家、省、市、区县，文本数据，长度不限。

5.2.13 家庭人均收入

研究对象家庭人口数/家庭结构、家庭人均收入、个人收入的数值，可是整数或浮点数，单位需明确，数值数据，长度不限。

5.3 观测指标和变量信息数据元

5.3.1 数据元命名

确保每个观测指标和变量的命名规范统一，避免歧义和混淆。采用标准命名规范有助于数据整合和后续分析。

5.3.2 数据元定义

提供每个观测指标和变量符合统一格式规范的结构化描述，包括定义、用途、时间或地点限定等信息，以便其他研究人员理解和正确使用数据。

5.3.3 数据类型和格式

描述每个观测指标和变量的数据类型（文本数据、数值数据等）和格式（例如日期的统一格式）。

5.3.4 值域和取值限制

定义每个观测指标和变量的可能取值范围，避免输入超出预期范围的值，确保数据的合法性和可信度。

5.3.5 缺失值表示方式

描述观测指标和变量的缺失值表示方式，例如可以选择特定的编码或标记。统一的缺失值表示有助于数据清洗和分析。

5.3.6 单位标准化

依据国际计量局（BIPM）制定和维护的《国际单位制》对每个观测指标和变量的单位标准化，避免不同来源的数据单位不一致，影响数据分析和比较的准确性。

5.4 数据采集信息数据元

5.4.1 数据采集日期

描述队列数据的采集日期，格式为YYYY-MM-DD。

5.4.2 数据采集方法

描述数据采集的方法，包括：问卷调查、临床检查、实验室检测、影像学检查、定性访谈、观察等。

5.4.3 数据采集工具

说明数据采集时使用的工具，包括：调查问卷、量表、病例报告表、诊断标准、实验室报告、影像学报告、访谈提纲等。

5.4.4 数据采集流程

描述数据采集的具体流程，包括采集前的准备、采集中的步骤、数据录入方式等。

5.4.5 质量控制措施及质量控制结果

说明数据采集时采用的质量控制措施，包括：数据核查、电话核查、录音核查、实地核查等。说明质量控制的标准并报告质量控制结果。

5.5 随访信息数据元

5.5.1 随访目的

说明随访的目的，明确随访时需要采集的信息和指标。

5.5.2 随访时间点和频率

描述随访的时间点和频率，包括开始和结束时间，以及随访实际接触的时间的间隔。

5.5.3 随访方式

描述主动随访的具体方式，如电话随访、面对面随访、电子邮件随访等，以及被动随访的具体方式，如死因监测、肿瘤登记、医院病案报告等。

5.5.4 随访结果及失访原因

说明随访的结果，包括完成随访和失访两类。对于失访，需注明原因，包括：拒访、无法联系、死亡，其中死亡如果作为随访结局指标则视为完成随访。

5.5.5 结局

研究随访过程中出现的预期的结果事件。可以是疾病发生，健康情况改变，或死亡等。

5.5.6 结局发生时间

在随访过程中观察到的结局出现的时间。

5.5.7 死亡原因

说明导致研究对象死亡的疾病的诊断代码，诊断代码参见GB/T 14396。

参 考 文 献

[1]GB/T 7408 数据元和交换格式 信息交换 日期和时间表示法
[2]T/CPMA 004-2019《大型人群队列终点事件长期随访技术规范》