**ICS** 11.020

|  |
| --- |
| **C** 07 |
|  |

团体标准

**T/BRA-CDCHE/ XXXX-XXXX**

|  |
| --- |
|  |

医学队列平台数据技术

队列数据管理 质控数据元定义及编码要求

Medical Cohort Platform Data Technology——

Cohort Data Managements——Definitions of Quality Control Element and Coding Verification

|  |
| --- |
| (草案) |
|  |

20XX- XX - XX发布

20XX - XX - XX实施

北京慢性病防治与健康教育研究会 发布

 目  次

[前  言 III](#_Toc168880263)

[1． 范围 1](#_Toc168880264)

[2． 规范性引用文件 1](#_Toc168880265)

[3． 术语和定义 1](#_Toc168880266)

[4． 质控数据元模型 2](#_Toc168880268)

[5． 数据元定义及编码 2](#_Toc168880269)

[5.1 队列数据基本信息数据元 2](#_Toc168880270)

[5.1.1 队列名称 2](#_Toc168880271)

[5.1.2 队列标识符 2](#_Toc168880272)

[5.1.3 研究目的 3](#_Toc168880273)

[5.1.4 研究背景 3](#_Toc168880274)

[5.1.5 研究范围 3](#_Toc168880275)

[5.1.6 研究样本量 3](#_Toc168880276)

[5.1.7 数据收集时间 3](#_Toc168880277)

[5.1.8 研究团队 3](#_Toc168880278)

[5.1.9 数据来源 3](#_Toc168880279)

[5.1.10 数据存储和管理 3](#_Toc168880280)

[5.1.11 隐私和伦理考虑 3](#_Toc168880281)

[5.2 队列数据研究对象信息数据元 3](#_Toc168880282)

[5.2.1 身份识别信息 3](#_Toc168880283)

[5.2.2 样本编码 3](#_Toc168880284)

[5.2.3 姓名 3](#_Toc168880285)

[5.2.4 出生日期 3](#_Toc168880286)

[5.2.5 年龄 3](#_Toc168880287)

[5.2.6 性别 4](#_Toc168880288)

[5.2.7 教育水平 4](#_Toc168880289)

[5.2.8 民族 4](#_Toc168880290)

[5.2.9 婚姻状况 4](#_Toc168880291)

[5.2.10 职业类别 4](#_Toc168880292)

[5.2.11 出生地 4](#_Toc168880293)

[5.2.12 现居地 4](#_Toc168880294)

[5.2.13 家庭人均收入 4](#_Toc168880295)

[5.3 观测指标和变量信息数据元 4](#_Toc168880296)

[5.3.1 数据元命名 4](#_Toc168880297)

[5.3.2 数据元定义 4](#_Toc168880298)

[5.3.3 数据类型和格式 4](#_Toc168880299)

[5.3.4 值域和取值限制 4](#_Toc168880300)

[5.3.5 缺失值的表示方式 5](#_Toc168880301)

[5.3.6 单位标准化 5](#_Toc168880302)

[5.4 数据采集信息数据元 5](#_Toc168880303)

[5.4.1 数据采集时间点 5](#_Toc168880304)

[5.4.2 数据采集方法 5](#_Toc168880305)

[5.4.3 数据采集工具 5](#_Toc168880306)

[5.4.4 数据采集流程 5](#_Toc168880307)

[5.4.5 质量控制措施及质量控制结果 5](#_Toc168880308)

[5.5 随访信息数据元 5](#_Toc168880309)

[5.5.1 随访目的 5](#_Toc168880310)

[5.5.2 随访时间点和频率 5](#_Toc168880311)

[5.5.3 随访方式 5](#_Toc168880312)

[5.5.4 随访结果及失访原因 5](#_Toc168880313)

[参考文献 6](#_Toc168880314)

 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。

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、中国医学科学院肿瘤医院、北京大学、山东大学齐鲁医院、北京大学第一医院、天津市安定医院、中国疾病预防控制研究中心、中国电子技术标准化研究院。

本文件主要起草人：刘肇瑞、魏文强、孙可欣、黄雨、岳伟华

本文件其它起草人：张婷婷、吕明、李航、徐广明、陈园生、李瑞琪、杨孝荣、陈浩、白倩倩、尹慧芳、李泊萱、王悦。

队列数据管理 质控数据元定义及编码

1. 范围

本文件规定了在开展跨队列研究时，原始队列数据集质控数据元定义及编码等内容。

本文件适用于指导拟开展跨队列研究的原始队列数据集的质量工作，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机构建立的人群队列。

1. 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 4761 家庭关系代码

GB/T 14396 疾病分类与代码

GB/T 7408 数据元和交换格式 信息交换 日期和时间表示法

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（系列标准）

WS 364 卫生信息数据元值域代码（系列标准）

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（系列标准）

1. 术语和定义

下列术语和定义适用于本文件。

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。

* 1.

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。

* 1.

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]

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]

1. 质控数据元模型

队列数据集的质控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5组数据元组成。

1. 数据元定义及编码
	1. 队列数据基本信息数据元

队列数据基本信息数据元应包含但不限于以下内容：

* + 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，但建议不超过40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。

* + 1. 队列标识符

队列数据的唯一标识，文本数据，由英文、数字或特殊字符在组成，长度不限。

* + 1. 研究目的

描述研究的目的、研究问题、研究假设等，文本数据，可为中文或英文，长度不限。

* + 1. 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。

* + 1. 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。

* + 1. 研究样本估计

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。

* + 1. 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。

* + 1. 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），数值信息，为10个字符。

* + 1. 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。

* + 1. 数据来源

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

* + 1. 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。

* + 1. 隐私和伦理考虑

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。

* 1. 队列数据研究对象信息数据元

为开展跨队列研究工作，队列数据研究对象信息数据元应包括但不限于以下内容：

* + 1. 身份识别信息

必须包括可以识别研究对象身份的信息，包括但不限于身份证号、医疗保险号、病历号等，文本数据，参照WS364.3 CV02.01.101 身份证件类别代码表、 GB 11643相关标准。

* + 1. 研究对象编号

必须包括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，文本数据，长度不限。

* + 1. 姓名

推荐包括研究对象本人在公安户籍管理部门正式登记注册的姓氏和名称，文本数据，最大20个字符。

* + 1. 出生日期

必须包括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。

* + 1. 年龄

必须包括研究对象参与队列数据采集时的周岁年龄，数值数据，为不超过3个字符的整数。

* + 1. 性别

必须包括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相关标准。

* + 1. 教育水平

必须包括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。

* + 1. 民族

必须包括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。

* + 1. 婚姻状况

必须包括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，最大2个字符，应符合GB/T 2261.2相关标准。

* + 1. 职业类别

必须包括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相关标准。

* + 1. 出生地

推荐包括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相关标准。

* + 1. 常住地

推荐包括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、街道/乡镇、详细地址等，文本数据，长度不限。

* + 1. 家庭人均收入

推荐包括研究对象家庭人均收入的数值，可以是整数或浮点数，单位需明确，数值数据，长度不限。

* 1. 观测指标和变量信息数据元

队列数据中观测指标和变量信息数据元属性应通过数据元专用属性表，即数据字典表示，包括但不限于以下内容：

* + 1. 数据元命名

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

* + 1. 数据元定义

提供每个观测指标和变量的描述，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。

* + 1. 数据类型和格式

描述每个观测指标和变量的数据类型（整数、浮点数、字符串、日期等）和格式（例如日期的统一格式）。

* + 1. 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。

* + 1. 缺失值的表示方式

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。

* + 1. 单位标准化

确保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。

* 1. 数据采集信息数据元

对队列数据的采集过程进行描述，包括但不限于以下内容：

* + 1. 数据采集时间点

明确队列数据的采集时间点或时间段，格式为YYYY-MM-DD，包括开始和结束时间，以确保其他研究者融合数据时选择跨度一致的数据库。

* + 1. 数据采集方法

描述数据采集的方法，包括但不限于：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。

* + 1. 数据采集工具

说明数据采集时使用的工具，包括但不限于：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。

* + 1. 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。

* + 1. 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，包括但不限于：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。

* 1. 随访信息数据元

当涉及重复多次采集数据时，需描述随访相关的信息，包括但不限于以下内容：

* + 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。

* + 1. 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，以及随访的间隔。

* + 1. 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。

* + 1. 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，包括但不限于：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。

参 考 文 献

1. Bernardi, F. A., D. Alves, N. Crepaldi, D. B. Yamada, V. C. Lima and R. Rijo (2023). "Data Quality in Health Research: Integrative Literature Review." J Med Internet Res 25: e41446.
2. Bhattacharya, A. A., E. Allen, N. Umar, A. Audu, H. Felix, J. Schellenberg and T. Marchant (2020). "Improving the quality of routine maternal and newborn data captured in primary health facilities in Gombe State, Northeastern Nigeria: a before-and-after study." BMJ Open 10(12): e038174.
3. Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly and J. Marchini (2018). "The UK Biobank resource with deep phenotyping and genomic data." Nature 562(7726): 203-209.
4. Cheikh Ismail, L., H. E. Knight, E. O. Ohuma, L. Hoch and W. C. Chumlea (2013). "Anthropometric standardisation and quality control protocols for the construction of new, international, fetal and newborn growth standards: the INTERGROWTH-21st Project." Bjog 120 Suppl 2(0 2): 48-55, v.
5. Damanabi, S., S. Abdolnejad and G. Karimi (2015). "Suggested Minimum Data Set for Speech Therapy Centers Affiliated to Tabriz University of Medical Sciences." Acta Inform Med 23(4): 243-247.
6. de Groot, R., D. P. Püttmann, L. M. Fleuren, P. J. Thoral, P. W. G. Elbers, N. F. de Keizer and R. Cornet (2023). "Determining and assessing characteristics of data element names impacting the performance of annotation using Usagi." Int J Med Inform 178: 105200.
7. de Lusignan, S., S. T. Liaw, P. Krause, V. Curcin, M. T. Vicente, G. Michalakidis, L. Agreus, P. Leysen, N. Shaw and K. Mendis (2011). "Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation. Contribution of the IMIA Primary Health Care Informatics Working Group." Yearb Med Inform 6: 112-120.
8. Downey, P. and T. C. Peakman (2008). "Design and implementation of a high-throughput biological sample processing facility using modern manufacturing principles." Int J Epidemiol 37 Suppl 1: i46-50.
9. Duda, S. N., B. E. Shepherd, C. S. Gadd, D. R. Masys and C. C. McGowan (2012). "Measuring the quality of observational study data in an international HIV research network." PLoS One 7(4): e33908.
10. Gallorini, M. (1995). "Standard reference materials and data quality assurance: the lesson from the analysis of trace elements." Toxicol Lett 77(1-3): 209-212.
11. Gallorini, M. and P. Apostoli (1996). "Standard reference materials and data quality assurance in the biomedical analysis of trace elements." Biol Trace Elem Res 52(3): 263-272.
12. Gassman, J. J., W. W. Owen, T. E. Kuntz, J. P. Martin and W. P. Amoroso (1995). "Data quality assurance, monitoring, and reporting." Control Clin Trials 16(2 Suppl): 104s-136s.
13. Hills, R. A., D. Revere, R. Altamore, N. F. Abernethy and W. B. Lober (2012). "Timeliness and data element completeness of immunization data in Washington State in 2010: a comparison of data exchange methods." AMIA Annu Symp Proc 2012: 340-349.
14. Huang, Y., Z. Liu, H. Wang, X. Guan, H. Chen, C. Ma, Q. Li, J. Yan, Y. Yu, C. Kou, X. Xu, J. Lu, Z. Wang, L. Liu, Y. Xu, Y. He, T. Li, W. Guo, H. Tian, G. Xu, X. Xu, S. Lv, L. Wang, L. Wang, Y. Yan, B. Wang, S. Xiao, L. Zhou, L. Li and L. Tan (2016). "The China Mental Health Survey (CMHS): I. background, aims and measures." Soc Psychiatry Psychiatr Epidemiol 51(11): 1559-1569.
15. Huang, Y., Y. Wang, H. Wang, Z. Liu, X. Yu, J. Yan, Y. Yu, C. Kou, X. Xu, J. Lu, Z. Wang, S. He, Y. Xu, Y. He, T. Li, W. Guo, H. Tian, G. Xu, X. Xu, Y. Ma, L. Wang, L. Wang, Y. Yan, B. Wang, S. Xiao, L. Zhou, L. Li, L. Tan, T. Zhang, C. Ma, Q. Li, H. Ding, H. Geng, F. Jia, J. Shi, S. Wang, N. Zhang, X. Du, X. Du and Y. Wu (2019). "Prevalence of mental disorders in China: a cross-sectional epidemiological study." Lancet Psychiatry 6(3): 211-224.
16. Jiang, G., R. C. Kiefer, L. V. Rasmussen, H. R. Solbrig, H. Mo, J. A. Pacheco, J. Xu, E. Montague, W. K. Thompson, J. C. Denny, C. G. Chute and J. Pathak (2016). "Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution." J Biomed Inform 62: 232-242.
17. Kässens, J. C., L. Wienbrandt and D. Ellinghaus (2021). "BIGwas: Single-command quality control and association testing for multi-cohort and biobank-scale GWAS/PheWAS data." Gigascience 10(6).
18. Li, M., Y. Huang, Z. Liu, R. Shen, H. Chen, C. Ma, T. Zhang, S. Li and M. Prince (2020). "The association between frailty and incidence of dementia in Beijing: findings from 10/66 dementia research group population-based cohort study." BMC Geriatr 20(1): 138.
19. Liu, Z., Y. Huang, P. Lv, T. Zhang, H. Wang, Q. Li, J. Yan, Y. Yu, C. Kou, X. Xu, J. Lu, Z. Wang, H. Qiu, Y. Xu, Y. He, T. Li, W. Guo, H. Tian, G. Xu, X. Xu, Y. Ma, L. Wang, L. Wang, Y. Yan, B. Wang, S. Xiao, L. Zhou, L. Li, L. Tan, H. Chen and C. Ma (2016). "The China Mental Health Survey: II. Design and field procedures." Soc Psychiatry Psychiatr Epidemiol 51(11): 1547-1557.
20. Parsa, N., M. J. Zibaeenezhad, M. Trevisan, A. Karimi Akhormeh and M. Sayadi (2020). "Magnitude of the Quality Assurance, Quality Control, and Testing in the Shiraz Cohort Heart Study." Biomed Res Int 2020: 8179795.
21. Peakman, T. C. and P. Elliott (2008). "The UK Biobank sample handling and storage validation studies." Int J Epidemiol 37 Suppl 1: i2-6.
22. Prorok, J. C., P. R. Williamson, B. Shea, D. Rolfson, L. R. Mañas, M. Cesari, P. Kim and J. Muscedere (2022). "An international Delphi consensus process to determine a common data element and core outcome set for frailty: FOCUS (The Frailty Outcomes Consensus Project)." BMC Geriatr 22(1): 284.
23. Rogers, J. R., T. J. Callahan, T. Kang, A. Bauck, R. Khare, J. S. Brown, M. G. Kahn and C. Weng (2019). "A Data Element-Function Conceptual Model for Data Quality Checks." EGEMS (Wash DC) 7(1): 17.
24. Stanfill, B. A., E. S. Nakayasu, L. M. Bramer, A. M. Thompson, C. K. Ansong, T. R. Clauss, M. A. Gritsenko, M. E. Monroe, R. J. Moore, D. J. Orton, P. D. Piehowski, A. A. Schepmoes, R. D. Smith, B. M. Webb-Robertson and T. O. Metz (2018). "Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data." Mol Cell Proteomics 17(9): 1824-1836.
25. Weiskopf, N. G. and C. Weng (2013). "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research." J Am Med Inform Assoc 20(1): 144-151.
26. Whitney, C. W., B. K. Lind and P. W. Wahl (1998). "Quality assurance and quality control in longitudinal studies." Epidemiol Rev 20(1): 71-80.
27. 冯牧青, 贺云, 高雪建, 严梅, 冯昌琪， 侯光辉 (2022). "数据集标识符与数据元内部标识符现状和应用研究." 医学信息学杂志 43(08): 47-51.
28. 金水高, 刘丽华, 王骏， 郭赁赟 (2006). "公共卫生信息系统数据元的标准化研究." 公共卫生与预防医学(01): 30-32.
29. 刘丹红, 王霞, 徐勇勇， 杨喆 (2012). "卫生信息标准化:从整理数据元到构建语义模型." 中国卫生信息管理杂志 9(04): 7-12.
30. 王霞 (2006). 卫生统计调查元数据概念模型的研究 博士.
31. 杨喆, 叶青, 刘丹红，徐勇勇 (2012). "医学观察数据的元数据框架及标准化应用." 中国数字医学 7(02): 37-40.
32. 尹榕慧, 姚祖发 (2020). "面向多领域标准的数据质量评估框架研究." 标准科学(01): 92-95.
33. 朱建平, 付学敬 (2023). "融合数据标准管理和数据血缘建设的数据治理方法." 上海质量(11): 49-55.