**团体标准编制说明**

《队列数据管理 质控数据元定义及编码要求》

(阶段：征求意见稿)

一、团体标准任务来源、编制背景、目的和意义

（一）任务来源

本标准根据2024年10月12日北京慢性病防治与健康教育研究会下达的《北京慢性病防治与健康教育研究会团体标准立项公告》，严格按照《医学队列平台数据技术》要求，由北京大学第六医院负责此项标准的起草制定工作。

（二）编制背景

《队列数据管理 质控数据元定义及编码要求》的编制背景主要是由于国内外在队列研究标准方面普遍缺乏适用于跨病种队列数据融合的质量标准体系。目前的研究环境中，缺乏针对跨病种队列数据融合前各单病种数据源的质量评价，也缺少对融合后形成的数据完整性以及应用的科学性等方面的系统评价。这导致各类专病及自然人群队列间的融合交叉研究在质量评价工作中面临困难。因此，编制这一标准旨在建立一个科学、系统的质量标准体系，以提升跨病种队列研究的整体质量和可信度。

（三）编制目的

《队列数据管理 质控数据元定义及编码要求》的编制目的在于建立一套系统化、标准化的数据元定义和编码规范，以确保队列数据的统一性和可比性。通过明确数据元的定义，能够有效减少数据在采集、存储和分析过程中的歧义和误差，提高数据的准确性和一致性。同时，编码要求的制定将促进不同队列间数据的互操作性，方便数据共享与融合，为跨病种研究提供坚实的数据基础。该标准的实施旨在提升数据质量管理水平，支持科学研究的严谨性和可重复性，最终推动队列研究成果的可靠性和有效性。

（三）编制意义

《队列数据管理 质控数据元定义及编码要求》的意义在于为队列研究提供清晰、规范的数据管理框架，确保数据的准确性、一致性和可比性。通过建立统一的数据元定义和编码标准，可以有效减少数据采集、存储和分析过程中的歧义与错误，从而提升数据质量。同时，该标准促进了不同队列间的数据互操作性，为数据共享和融合奠定了基础，支持跨病种研究的开展。此外，规范的数据管理有助于提高科学研究的透明度和可信度，为研究人员提供清晰的操作指导，最终推动科学研究的创新与发展，提升研究成果的社会和学术价值。

二、工作简况

（一）起草单位及主要起草人

本标准由北京大学第六医院主要负责起草，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参与，专家团队涵盖了精神、神经、肿瘤、消化、心血管、疾病监测以及自然人群等领域。

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 专业 | 职称 | 工作单位 |
| 刘肇瑞 | 精神障碍  流行病学 | 副研究员 | 北京大学第六医院 |
| 魏文强 | 流行病学 | 研究员 | 中国医学科学院肿瘤医院 |
| 孙可欣 | 流行病学 | 副研究员 | 中国医学科学院肿瘤医院 |
| 黄雨 | 计算机软件与理论 | 研究员 | 北京大学 |
| 岳伟华 | 精神病与精神卫生学 | 教授 | 北京大学第六医院 |
| 张婷婷 | 社会精神病学 | 主管技师 | 北京大学第六医院 |
| 吕明 | 流行病学 | 教授 | 山东大学齐鲁医院 |
| 李航 | 皮肤病与图像学 | 主任医师、教授 | 北京大学第一医院 |
| 徐广明 | 社会精神病学 | 主任医师 | 天津市安定医院 |
| 陈园生 | 公共卫生 | 研究员 | 中国疾病预防控制中心 |
| 李瑞琪 | 控制科学与工程 | 高级工程师 | 中国电子技术标准化研究院 |
| 杨孝荣 | 循证医学 | 副研究员 | 山东大学齐鲁医院 |
| 陈浩 | 流行病学 | 助理研究员 | 山东大学齐鲁医院 |
| 白倩倩 | 皮肤病与性病学 | 主治医师 | 北京大学第一医院 |
| 尹慧芳 | 精神病与精神卫生学 | 副主任医师 | 天津市安定医院 |
| 李泊萱 | 精神病学 | 主治医师 | 天津市安定医院 |
| 王悦 | 公共卫生 | 助理研究员 | 中国疾病预防控制中心 |

（二）主要编制过程

2024年6月，北京大学第六医院同天津市安定医院、北京大学、中国医学科学院肿瘤医院、山东大学齐鲁医院、北京大学第一医院、中国电子技术标准化研究院、中国疾病预防控制研究中心等有关起草单位启动了标准前期研究项目“跨队列研究数据融合质量控制标准研发”,对跨队列研究数据融合工作的流程和质控环节展开调研，并撰写了调研报告，明确了跨队列研究数据融合工作的质量控制应基于工作流程开展，包括原始队列数据采集、原始队列数据元定义、队列数据融合三个步骤，同时对于各环节的质量控制报告应有统一要求。

2024年7月，北京大学第六医院会同其他起草单位开展准备团体标准立项相关材料工作，一同针对制定《队列数据管理 质控数据元定义及编码要求》团体标准的具体工作进行了部署，确定了总体工作方案，并组建了标准起草小组。

2023年8月-2024年9月，标准编写小组按照标准的修订程序进行标准编制，形成标准初稿，并在起草单位内部多次讨论和修改，形成了第一版的标准初稿，提交北京慢性病防治与健康教育研究会进行立项审查。

2024年9月，根据北京慢性病防治与健康教育研究会发布的“北京慢性病防治与健康教育研究会团体标准立项公告”，北京大学第六医院同其他起草单位向北京慢性病防治与健康教育研究会提交申请，明确了《队列数据管理 质控数据元定义及编码要求》编写工作组的具体成员，并确定了参编单位和参编人员排序。

2024年10月，北京慢性病防治与健康教育研究会同意该标准立项。立项下达后，北京大学第六医院同其他起草单位进一步对第一版《队列数据管理 质控数据元定义及编码要求》标准初稿进行讨论和修改，同时修订的还包括北京慢性病防治与健康教育研究会关于队列数据管理中的其他两项标准。最终形成了标准初稿，并提交北京慢性病防治与健康教育研究会进行初稿审核。

三、标准编制原则

（一）编制原则

1.科学规范。本标准按照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》中的规范性要求进行编写，对本文件规定了跨队列研究实施过程中质控数据元定义及核查质量控制等内容，进行规范性阐述。

2.系统全面。跨队列研究实施过程中本文件规定了跨队列研究实施过程中质控数据元定义及核查质量控制的全过程进行系统梳理，标准内容涵盖对数据元模型的定义和核查的质量控制处理等。对跨队列研究队列数据管理的数据融合质量控制进行全面规范，编制高水平的团体标准。

3.实用好用。尽可能从实用的角度出发，充分考虑标准的可实施性和实施效果，吸引了各方工作人员参与，多方面征集标准相关部门以及其他相关方的意见，完善标准内容。

4.协调一致。标准编制过程中，对相关的国家标准以及政策文件进行了充分的分析和考虑，保证本标准与国家、行业标准及政策文件的协调。

四、主要技术内容的论据

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核查等内容。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核查工作，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。本文件核心技术内容包括：质控数据元模型和数据元定义及核查等章节。

术语和定义：规定了队列、跨队列定义，结合实际经验，重点参考了《Connecting Cohorts to Diminish Alzheimer’s Disease (CONCORD-AD): A Report of an International Research Collaboration Network》和《Using a cross-cohort comparison design to test the role of maternal smoking in pregnancy in child mental health and learning: evidence from two UK cohorts born four decades apart》两篇文章中的相关论述。规定了数据元、属性、字符定义，来源于GB/T 18391.1-2009,3.3.8、GB/T 18391.1-2009,3.1.1和GB 18030-2022。

质控数据元模型：明确了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。

数据元定义及核查：分别对五组数据元进行说明。对于队列数据基本信息数据元，根据《中国精神卫生调查研究方案》和《临床队列遗传及共享平台建立研究手册》，结合实践经验，明确了能够描述队列数据基本信息的数据元，包括队列名称、队列标识符、研究目的、研究背景、研究范围、研究样本量、数据收集时间、研究团队、数据来源、数据存储和管理以及隐私和伦理考虑。对于队列数据研究对象信息数据元，规定了研究对象信息数据元的规则，数据编码的内容应符合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、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）。对于观测指标和变量信息数据元，结合实际经验，参考中华预防医学会T/CPMA 001-2018《大型人群队列研究数据处理技术规范》、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），对数据元命名、数据元定义、数据类型和格式、缺失值的标识方法、单位标准化进行了规定。对于数据采集信息数据元和随访信息数据元，根据数据管理实际，参考《中国精神卫生调查研究方案》和《临床队列遗传及共享平台建立研究手册》，规定了数据采集时间点、数据采集方法、数据采集工具、数据采集流程、质量控制措施及质量控制结果、随访目的、随访时间点和频率、随访方式、随访结果及失访原因。

五、主要试验、验证及试行结果

该项标准依托科技部国家重点研发计划《大型队列间联合研究柔性化大数据云平台支撑系统研发》项目课题一《跨队列研究及数据共享标准、机制及质量控制方法研究》以及2023年中国疾控中心公共卫生领域卫生健康标准化前期研究项目《跨队列研究数据融合质量控制标准研发前期研究项目》,对制定的标准进行试验方法验证。

在拥有存储10万余例重大疾病队列的结构化临床、随访和影像数据的临床科研数据中心以及近4万的社区人群队列进行了广泛的使用，并进行了大量的验证试验和处置工作，在此基础上对标准草案进行了论证、补充、细化和完善。目前已形成了一套稳定的数据采集流程和数据质量控制方法，基本上能够保证数据质量，促进数据融合和分析。

六、采用国际标准或国外先进标准的情况

美国食品药品监督管理局（FDA）发布的“STUDY DATA TECHNICAL CONFORMANCE GUIDE”（研究数据结构化技术指南），该指南对提交至FDA动物和人体研究数据提出了标准化规范，以便于FDA对于科研项目的评审。但没有对队列数据的质控数据元进行定义及核查的标准。

七、与现行法律法规、强制性标准和其他有关标准的关系

1、法律法规依据

《国家健康医疗大数据标准、安全和服务管理办法（试行）》第九条 国家卫生健康委员会鼓励医疗卫生机构、科研教育单位、相关企业或行业协会、社会团体等参与健康医疗大数据标准制定工作。公民、法人或者其他组织可提出制修订健康医疗大数据标准的立项建议，并提交相应标准项目建议书。第三十三条 责任单位应当结合服务和管理工作需要，及时更新、甄别、优化和维护健康医疗大数据，确保信息处于最新、连续、有效、优质和安全状态。

2、相关标准

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

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）

八、重大意见分歧的处理经过和依据

无

九、作为推荐性标准的建议

无

十、贯彻该标准的要求、措施建议及预期效果

在现代科学研究中，跨队列研究作为一种重要的研究手段，面临着数据一致性、共享和管理效率等诸多挑战。因此，贯彻该标准的要求是提升跨队列研究质量的关键。首先，必须规范数据的定义和编码，确保各队列研究在数据采集、管理和分析中的一致性和可靠性。通过建立统一的质控数据元定义和核查标准，可以有效减少数据在不同队列间的误差与不一致，增强研究结果的科学性和可信度。

为实现这些目标，建议采取以下措施：一是制定详细的实施指南，明确数据采集、录入、校验和分析的标准流程；二是组织定期培训，提高研究人员对标准的理解和应用能力；三是建立质量监控机制，定期对数据质量进行评估，及时发现和纠正问题；四是推动数据共享平台的建设，促进不同队列间的数据互操作性，增强数据融合的能力。

预期效果方面，贯彻该标准将显著提升跨队列研究的数据质量和管理水平，促进数据共享与应用，提高科学研究的创新能力。同时，统一的数据管理标准将简化管理流程，减少重复劳动，提高数据处理和分析的效率，为研究人员节约时间和成本。此外，遵循国际通用标准，将使得研究成果更易被国际学术界接受和认可，促进国际交流与合作，提高我国在全球科研领域的影响力。

总体而言，贯彻《队列数据管理 质控数据元定义及编码要求》标准将为跨病种队列数据分析提供坚实的保障，助力科学研究在不同时期采集的队列数据中充分体现其价值，推动科学研究的持续发展。

十一、废止或代替现行相关标准的建议

无

十二、其他应予说明的事项

无