ICS 11.020
CCS C07

团 体 标 准

T/BRACDCHE 006-2025

跨队列研究数据融合质量控制要求

Quality Control Requirements of Data Fusion 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　融合准备 1
4.1　融合的目的 1
4.2　队列数据集概述 1
4.3　数据可用性评估 2
5　融合中的规范 3
5.1　清理拟融合的原始队列数据 3
5.2　数据融合规则 4
5.3　根据设定的数据融合标准评估所需的变量并进行定量评分 5
5.4　验证融合结果 5
6　融合后的要求 5
6.1　过程记录 5
6.2　确保过程的可追溯性和透明度 5
6.3　定期评估和更新数据集 5
7　数据安全要求 5

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、中国医学科学院肿瘤医院、北京大学、中国疾病预防控制中心、山东大学齐鲁医院、天津市安定医院、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、张婷婷、魏文强、孙可欣、陈冬雪、黄雨、陈园生、丁若溪、罗雅楠、吕明、徐广明、李明慧、黄悦勤、邓咏妍、张媛、李航、尹慧芳、李瑞琪、王悦、张同超、白倩倩、葛红敏、潘鹏、颜国利。

跨队列研究数据融合质量控制要求

1 范围

本文件规定了跨队列研究实施过程中数据融合的质控要求等内容。

本文件适用于指导拟开展跨队列研究的数据融合工作，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。

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。

WS/T306-2023 卫生健康信息数据集分类与编码规则
WS 363-2011 卫生信息数据元目录（系列标准）
WS 364-2011 卫生信息数据元值域代码（系列标准）
WS 365-2011 城乡居民健康档案基本数据集
WS/T 370-2022 卫生健康信息基本数据集编制标准
WS 372-2012 疾病管理基本数据集
WS 375-2012 疾病控制基本数据集（系列标准）
GB/T 37973-2019 信息安全技术 大数据安全管理指南
GB/T 39725-2020 信息安全技术 健康医疗数据安全指南

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。

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
数据融合 data fusion
将来自不同来源和不同格式的数据进行整合和合并。

3.4
变量 variable
数据集中最基本的元素，用于存储数据。

4 融合准备

4.1 融合的目的

明确说明开展数据融合的原因和目标，并结合具体的研究背景，阐明数据融合如何提升研究的代表性、广泛性或统计效能。

4.2 队列数据集概述

4.2.1 队列数据集的基本信息

根据数据的可靠性、适用性和完整性，结合数据融合的目的，确定拟融合的队列数据集，并收集相关信息，包括数据来源和研究起止日期。包括数据来源的多样性和可靠性评估，如数据来源单位、研究起止日期、负责人、经费、数据采集的起始和完成日期等。

4.2.2 研究目的

明确队列数据的研究目的，明确研究内容与预期结果。

4.2.3 研究背景

说明数据融合涉及的研究背景，包括研究领域、研究现状以及相关的科学或社会意义，概述数据融合在解决当前研究问题中的作用和必要性。

4.2.4 研究设计

明确队列数据的研究设计类型，包括但不限于：现况调查、诊断试验、病例对照研究、队列研究、纵向研究、实验性研究等。对于每种研究设计类型，给出相应的定义，以明确其特征、方法以及类型之间的区别和联系。根据不同的研究设计类型，记录相应的设计要素，包括抽样方法、随机方法、暴露因素、观测指标、干预措施、结局指标等内容。

4.2.5 研究对象

明确研究对象的人口学、表现等信息特征、纳入标准、排除标准以及随访的期限。随访期限应根据研究设计类型、研究目标及数据的可获取性进行合理设定。通常，随访期限应不低于数据收集的最短周期，并与研究假设相一致。

4.2.6 变量

明确队列数据中的变量定义、数据类型、表示格式、数据单位以及取值范围。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）。

4.2.7 样本量

根据研究目的选择样本，评估样本量是否足够满足研究要求，包括统计学检验、结果精度和效能分析。确保样本量的选择符合研究设计的总体目标。

4.3 数据可用性评估

根据队列数据集的数据特征，评估拟进行数据融合的队列数据的可用性，评估内容包括但不限于数据清单、数据质量、核心变量和其他变量、队列研究信息、知情同意。

4.3.1 数据清单

创建一个清单，列出所有拟进行融合数据的队列数据源，包括其来源、格式、结构和可用性，从而对数据的整体情况进行全面了解。

4.3.2 缺失数据

对每个数据源进行缺失数据的识别和分析。确定每个数据源中缺失数据的变量数量、比例以及缺失原因。原因包括但不限于记录不完整、人为错误、技术问题或其他因素。

4.3.3 核心变量和其他变量

在研究过程中，应明确区分核心变量与其他变量：

a) 评估变量的重要性：评估每个变量对研究问题的贡献，分析其是否为主要观测指标，明确哪些变量对研究目标至关重要，哪些为次要变量。
b) 核心变量与其他变量的分类：确定核心变量和辅助变量，核心变量直接影响主要结果，辅助变量对研究起到补充作用。

c) 缺失数据处理的决策依据：明确变量缺失比例，并根据变量重要性权衡缺失数据处理方法，优先处理核心变量。
d) 数据合法使用和合规性：确保在使用队列数据之前获得使用许可，遵循法律、伦理和知情同意要求，确保数据合法合规使用。

4.3.4 数据开放情况或数据获取方式

明确数据的开放情况或获取方式，描述数据是否免费开放、是否需要付费，或者需要联系数据负责人获取数据。若数据免费开放，应说明数据可以在公开平台上访问和使用；若数据需要付费获取，应明确付费标准、支付方式以及相关条款；如数据需要通过联系负责人获取，应明确如何与数据负责人建立联系，包括联系方式及获取许可的流程。

4.3.5 提供跨队列研究信息

向数据负责人提供跨队列研究的详细信息，包括研究问题、研究目的、数据提取需求、数据融合统计分析计划等。

4.3.6 请求正式同意

请求数据负责人提供正式的知情同意文件，明确数据负责人同意跨队列研究使用、分享和融合他们的数据，提供者是请求数据人，签署者是数据负责人。在请求使用队列数据时，需要确认所获得的数据是否包含可识别个人身份的信息。如果数据中包含可识别的个人信息，应重新获得知情同意，明确告知数据的使用范围、目的、保密措施等内容，确保参与者的隐私权和数据安全得到充分保障。在知情同意过程中，需明确认定数据的持有权、使用权和经营权，以确保数据权益得到充分保护和尊重。知情同意文件可以是书面协议、合同或许可证书，获取后按照有效期对知情同意文件进行档案保存，知情同意文件包括但不限于以下内容：

a) 数据使用目的和范围；
b) 数据分享和融合的具体条款和条件；
c) 数据的保护和隐私保护措施；
d) 同意的有效期限，以及期限终止或数据不再使用时的处理措施。应明确数据使用期满后，如何处理数据，如销毁、匿名化或归还给原数据提供方；
e) 同意的签署日期和签名。

5 融合中的规范

5.1 清理拟融合的原始队列数据

5.1.1 统一变量的赋值方式

在进行数据融合时，需明确统一变量赋值的规则和方法，具体包括以下：

a) 制定融合规则和方法：在数据融合前，需制定详细且明确的融合规则和方法，确保核心变量和次要变量的选择符合研究需求。
b) 变量选择的评估标准：根据变量的研究重要性、数据质量及可用性，对变量进行合理评估和排序，确保核心变量能够最大程度地支持研究目标。
c) 核心变量与次要变量的功能区分：确保核心变量在数据融合中能直接支持研究目标，次要变量则应起到辅助作用，确保分析结果的完整性和准确性。
d) 统一变量赋值的考虑因素：在进行统一变量赋值时，需考虑变量的一致性和转换规则，以确保各个数据源之间的兼容性。
e) 处理单位不一致和编码差异：对于因单位不一致或赋值编码不一致导致的队列数据间差异，可通过数据转化方式进行统一。
f) 处理调查工具不一致的差异：对于调查工具不一致导致的差异，应采用标准化转换、映射关系建立或模型调整等方法进行统一。

5.1.2 数据差异转化与问题识别

针对因单位、编码或调查工具不一致导致的数据差异，通过建立统一转化规则、映射关系或模型实现统一，并识别原始数据问题和原始数据的质量问题，如缺失值、异常值、重复值等。

5.1.3 处理缺失数据

可根据研究需求决定在数据融合过程中是否对数据的缺失值进行处理，包括以下内容：

a) 删除缺失数据行：缺失数据占比不超过 10%的情况下，可以删除缺失数据行
b) 插补数据：对于重要的核心变量，缺失比例超过 10%时，可考虑使用统计插补方法（如均值插补、回归插补等）；
c) 分析缺失数据：对于某些数据缺失原因明确，且缺失本身可能提供额外信息时，可将缺失数据作为独立变量进行分析；
d) 加权分析：当缺失数据占比较大时，可通过加权分析调整样本，减少缺失对结果的偏差影响。

5.1.4 处理异常数据

可根据研究需求决定在数据融合过程中是否对数据的异常值进行处理，根据异常值的特征和对数据的理解，选择适当的处理策略，包括以下内容：

a) 删除异常值：若缺失变量过多且影响显著，可删除整行数据；若缺失变量较少且在可接受范围内，应优先采用数据插补以保留数据完整性；
b) 修复异常值：可通过插值、取中位数、平均值等方式对异常值进行修复，使修复后的异常值更接近数据集的整体趋势；
c) 分组：将异常值分到适当的分组中，以降低其对分析的影响；
d) 转换数据：应用适当的数学函数对异常值进行转换，降低其对数据分析的影响。

5.1.5 处理重复数据

根据重复数据的特征和对数据的理解，选择适当的处理策略，包括以下内容：

重复记录：

a) 删除重复记录：删除完全重复的记录，仅保留一条唯一记录；
b) 合并重复记录：将重复的记录合并为一条唯一记录，保留合并后的数据；
c) 分析重复记录原因：对不明确原因的重复记录进行分析，修复数据采集或输入过程中的错误。

重复变量：

a) 删除重复数据：删除完全重复的变量，保留一个变量；
b) 合并重复数据：对于表达相同信息的重复变量，考虑合并为一个变量；
c) 分析重复数据原因：对重复变量的产生原因进行分析，检查数据收集或变量定义中的潜在问题。

5.1.6 与研究目标的权衡

在决定如何处理缺失数据和异常值时，需要权衡数据的可用性和对研究目标的影响。如果数据的缺失和异常值对研究问题非常关键，那么需要采取更谨慎的方法来处理缺失数据，如插补。如果数据的缺失和异常值对研究问题影响较小，可考虑简化的方法，如删除缺失数据行。

5.1.7 记录数据清理过程

记录原始数据的过程，使用的数据清理工具、脚本或软件，并记录其参数和设置，缺失值、异常值、重复值的处理方式，以及数据转换，变量的删除或保留及其原因。使用的工具和方法应符合行业标准，并具有可追溯性和一致性。

5.1.8 数据质量验证

对清理后的数据质量进行验证，确保数据质量满足要求。

5.2 数据融合规则

制定明确的融合规则和方法，包括数据匹配方式、变量匹配、合并方式等。详细说明数据匹配方法，如使用精确匹配、模糊匹配或基于算法的匹配方式（例如距离度量、相似度分析等），以确保不同数据源间的正确对接和融合。

5.2.1 选择需要融合的核心变量和次要变量

根据研究目的和问题，选择需要融合的核心变量和其他变量，并评估这些变量的重要性及局限性。

5.2.2 设定数据质量评估指标

评估指标应包括数据准确性、完整性、一致性和可靠性。建议采用某种评分标准或使用具体的数据质量检测工具，以便于对数据质量进行量化评估，例如：使用0-100分的评分标准对数据准确性、完整性、一致性和可靠性进行量化评估，或引用相关数据质量检测工具（如数据质量评估框架）进行数据质量评估。

5.2.3 设定预期满足标准的数据变量的百分比

设定预期满足标准的数据变量的百分比是为了确保在融合过程中，数据质量达到一定的标准。这个百分比应基于评估数据质量的指标和方法来设定。

5.3 根据设定的数据融合标准评估所需的变量并进行定量评分

根据设定的数据融合标准，对所需的变量进行评估，并根据预设标准进行定量评分。这些评分可帮助在融合过程中权衡不同变量的重要性和可靠性。

5.4 验证融合结果

验证融合后数据集的准确性和一致性，确保融合结果符合预期。

6 融合后的要求

6.1 过程记录

记录数据融合的所有步骤、规则、参数以及处理过程，以便后续审查和复制研究。

6.2 确保过程的可追溯性和透明度

确保数据融合的过程具有可追溯性，能够清晰展示每个数据源如何被融合到最终结果中。

6.3 定期评估和更新数据集

定期评估数据融合的效果，根据反馈和新的需求更新融合标准，以保持数据融合的质量和适应性。

7 数据安全要求

大型队列数据融合过程中的安全管理应遵循《GB/T 37973-2019 信息安全技术 大数据安全管理指南》、《GB/T 39725-2020 信息安全技术 健康医疗数据安全指南》中的相关要求。应加强数据融合过程的透明度和可追溯性，具体要求包括：明确数据融合过程中的关键决策点，记录数据变更情况，并建立数据融合结果的审计追踪机制，以确保数据处理的完整性和透明度。