标准·方案·指南

大型人群队列研究数据处理技术规范（T/CPMA 001–2018）

中华预防医学会
通信作者：李立明，Email：lmleeph@vip.163.com
基金项目：国家重点研发计划（2016YFC0900500，2016YFC0900504）
DOI：10.3760/cma.j.issn.0254–6450.2019.01.003

Technical specification of data processing for large population-based cohort study（T/CPMA 001–2018）

Chinese Preventive Medicine Association
Corresponding author: Li Liming, Email: lmleeph@vip.163.com
Fund programs: National Key Research and Development Program (2016YFC0900500, 2016YFC0900504)
DOI：10.3760/cma.j.issn.0254–6450.2019.01.003

前　言

本标准按照GB/T 1.1—2009给出的规则起草。

本标准由北京大学提出，中华预防医学会归口。

主要起草单位：北京大学、中国医学科学院、北京理工大学。

本标准主要起草人：李立明、余灿清、吕筠、卞铮、谭云龙、刘亚宁、郭彧、汤海京、杨旭。

本标准为首次发布。

引　言

大型人群队列研究数据内容丰富、来源多样，规范而准确的数据是高质量队列研究的基本要求之一。

大型人群队列研究数据的管理和利用应遵循一定的原则和规范，依次进行数据标准化、清理及质控和数据整合。数据标准化应当遵循系统性、科学性、统一性和可用性的原则，从数据处理计划开始，涉及数据类型、格式、值、衍生和编码等多个方面。经数据标准化后，还应进行数据清理和质控，对数据进行全面的检查并给予相应的处置，保证数据达到规范性、完整性和准确性等质量要求。由于队列研究数据来源多样，最后应整合到项目的标准化数据库中。在数据整合过程中，应综合考虑数据来源、数据特征等方面的因素，确保实现队列数据的高效存储和利用。

大型人群队列研究数据处理技术规范

1　范围

本标准规定了大型人群队列研究实施过程中数据标准化、清理、质控及整合的基本原则。

本标准对不同来源、不同类型的队列数据标准化、清理、质控及整合进行规范化要求，适用于已建立或拟开展大型人群队列研究的机构，包括但不限于大型自然人群队列、区域性人群队列、针对某一特种疾病或基于特殊机构开展的人群队列。本标准还可供规模相对较小的人群队列研究参考。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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 1.1 标准化工作导则 第1部分：标准的结构和编写规则（GB/T 1.1—2009，ISO/IEC Directives，Part 2，2004）

3　术语和定义

下列术语和定义适用于本文件。

3.1　队列研究 cohort study

队列研究是将一个范围明确的人群按是否暴露于某可疑因素或其暴露程度分为不同的亚组，追踪其各自的结局，比较不同亚组之间结局的差异，从而判定暴露因子与结局之间有无因果关联及关联大小的一种观察性研究方法。

3.2　数据 data

数据是指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合。它不仅指狭义上的数字，还可以是具有一定意义的文字、字母、数字符号的组合、图形、图像、视频、音频等，也是客

观事物的属性、数量、位置及其相互关系的抽象表示。

3.3 数据库 database

数据库，或称电子数据库，是指按照数据结构来组织、存储和管理数据的仓库，它是以一定方式存储在一起、能为多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。

3.4 数据集 dataset

数据集是指数据的集合。最常见的形式是数据表，其中每一列代表一个变量，每一行代表一个观察记录。

3.5 数据标准化 data standardization

数据标准化是指将数据转换成某种统一形式的过程。

3.6 编码字典 codebook

编码字典，或称值域代码，是指记录编码及其相应属性的文件。

3.7 数据清理 data cleaning

数据清理是指对数据进行重新审查和校验，发现并纠正数据文件中可识别错误的过程。

3.8 研究对象 study subject

研究对象是指样本人群中符合纳入和排除标准的合格对象。

3.9 个体唯一性标识 personal unique identification

个体唯一性标识是指每一名研究对象特有的，可以唯一识别其自然人个人身份的信息，包括身份证号码、医疗或社会保险号码等。

3.10 常规监测 routine surveillance

常规监测是指通过相关政府部门（包括卫生、公安、民政、社会保障、计划生育等）当前运行的各类监测系统或常规工作中形成的资料和数据库，从中筛选出研究所需的随访信息，收集研究对象各类死亡、发病、迁移和失访等终点事件。

3.11 社区定向监测 community targeted surveillance

社区定向监测是指将研究对象的名单提供给研究社区街道、居委会或乡镇、村的相关工作人员，定期联系研究对象，从而获取该社区内研究对象的死亡、发病、迁移等有关随访信息。

3.12 失访 loss to follow up

失访是指队列研究中，户口已迁出调查区域，且经查找仍无法得知去向，或虽有明确下落，但无法进行长期随访监测（如户口搬迁到外地等）的研究对象。

3.13 非结构化数据 unstructured data

非结构化数据是指数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据。包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等。

3.14 数据整合 data consolidation

数据整合是指将不同数据源的数据收集、整理、清洗，转换后加载到一个新的数据源，为数据消费者/数据使用者提供统一数据视图的数据集成方式。

3.15 数据脱敏 data masking

数据脱敏是指利用随机字符或数字加密隐藏原始数据的过程。一般需要进行脱敏处理的数据包括个人识别数据、个人敏感数据等。

4 数据标准化

4.1 基本要求

对研究数据进行标准化的目的，是为了保证数据集内部的一致（consistency），也为了便于数据集间的整合。对数据的标准化处理应满足如下要求：

4.1.1 一致性

即数据集或数据库内部的标准（如：变量定义、格式、单位、取值精度、编码规则等）应保持一致。

4.1.2 通用性

即数据与其他外部数据的标准应尽量保持一致，宜参考或使用现行或通用的卫生相关数据集标准，尤其是需要与外部数据进行链接时。

4.1.3 易用性

即标准化之后的数据应尽量清晰易懂，并且方便进行进一步的数据清理、整合与分析。

4.2 实施过程

4.2.1 数据标准化之前，应制定详细的数据处理计划，其中应包括：

a）原始数据的来源、性质、内容；

b）数据库的设计方案；

c）准备处理的文件和变量，以及相应的标准化处理方案；

d）准备生成的新变量和生成方法；

e）准备予以编码的变量，以及编码方式。

4.2.2 按照数据处理方案，对数据文件进行标准化处理，并且详细记录每一步的处理方法与结果。

4.2.3 数据处理完成后，应准备详细的说明文件，对标准化之后的数据予以必要的说明与解释。

4.2.4 数据处理过程中，应尽量保存原始数据或每一个步骤的中间数据，以备回顾和检查。

4.3 数据库设计

4.3.1 当研究使用关系型数据库来储存数据时，应在数据收集之前设计数据库。

4.3.2 数据库设计既要满足研究需要，又要尽量做到精简、避免重复。设计应符合关系型数据库的第三范式，基本要求包括：

a）将不同种类的数据存放于不同的位置，如基

线调查数据与随访数据；

b）数据之间能够建立关联；

c）不重复存放冗余的数据；

d）命名清晰易懂，并且保持一致。

4.4 数据类型标准化

应将收集到的数据（变量）设置为适当的类型。此类数据通常为结构化数据，类型一般包括数值、字符串和日期/时间3种，适用于标准化的数据分别是：

4.4.1 数值型

4.4.1.1 数值型适用于各类计量的变量，例如定量的检查指标、计数的项目等。

4.4.1.2 数值型变量可进一步按照是否保留小数位数，分为整数型和小数型两类，其适用的数据如下：

a）整数型：适用于计数的项目，例如子女的个数；

b）小数型：适用于精确度要求较高，需要保留小数位数的项目，如体重；或通过对整数的计算而生成项目，如体质指数。

4.4.1.3 对于一些将定性项目分类编码后的变量，出于易用性的考虑，可设置成数值型。例如，将男性和女性分别编码为0和1后，该编码可设置成数值型，并在编码字典中设置相应的标签。具体参见4.7部分内容。

4.4.2 字符串型

4.4.2.1 字符串型适用于除定量项目外，各类文字描述，或定性表示的变量，例如姓名、地址等。

4.4.2.2 对于一些将定性项目分类编码后的变量，除外如5.1.3所指的情况外，应设置为字符串。例如，将全国的省份分别编码之后，该编码应设置成字符串。

4.4.3 日期/时间型

4.4.3.1 日期/时间型适用于所有表示日期或时间的变量，如出生日期、检查时间等。

4.4.3.2 日期/时间型可进一步按照是否保留日期和时间这2个成分，分为日期型、时间型和日期/时间型3类，其适用的数据分别是：

a）日期型：适用于仅需要考虑日期，不需考虑时间的变量；

b）时间型：适用于仅需考虑时间，不需考虑日期的变量；

c）日期/时间型：适用于需要同时考虑日期和时间的变量。

4.5 数据格式和值的标准化

根据研究需要，将数据的格式和值设置为统一的格式。

4.5.1 对于有单位的计量变量，应将取值转化成通用单位下的值。例如质量统一转化为公斤，长度统一转化为米。

4.5.2 对于小数型变量，应将取值转化为统一的小数位数。

4.5.3 对于日期/时间变量，应转化成统一的格式，例如YYYY/MM/DD HH:MM:SS。

4.5.4 对于文本型的数据，宜使用统一的术语与形式。例如地址，宜统一转化为“XX省XX市XX区XX街XX号”的形式；有多种名称的疾病，采用统一的名称。

4.6 用标准方式生成新变量

对于需要通过计算而生成的新变量，应采用标准或通用的方式或公式。例如，对于体质指数，其计算方法是体重（kg）除以身高（m）的平方。

4.7 标准编码

4.7.1 对于分类变量，宜予以编码，即用号码来代表相应的类别。

4.7.2 编码方法应保持一致。可自行制定编码规则和方法，也可采用一些通用的标准编码。例如，对于疾病，可采用国际疾病分类（International Classification of Diseases）进行编码。

4.7.3 编码完成后，宜设置相应的值标签，或者建立编码字典。

5 数据清理及质控

大型人群队列研究的数据主要来源于现场调查和长期随访监测。现场调查可综合采用问卷调查、体格测量、生物样本采集等方法收集暴露数据。长期随访监测可通过重复调查、常规监测和社区定向监测等方法获取结局数据。这种多来源的数据经标准化后，还应进行数据清理及质控，保证数据符合规范性、完整性和准确性等质量要求。数据清理及质控流程可分为数据检查、问题处置和统计学监测等环节。

5.1 数据检查

5.1.1 规范性核查

应对数据文件和变量属性进行规范性评价，核查其是否符合现行的或本研究制定的数据标准、规范或要求。若数据已经过标准化处理，则可省略此步骤。

5.1.2 完整性核查

应对数据集的样本量和变量信息进行完整性评价，识别缺失数据。缺失类型如下：

5.1.2.1 记录缺失。除外重复数据，应核查数据集的实际样本量或记录数与应获取数目是否相同。

5.1.2.2 变量缺失。除外重复变量，应核查数据集中已有变量数是否少于应获取的变量数。

5.1.2.3 变量值缺失。应核查数据集中特定单元格是否存在信息缺失。

5.1.3 唯一性核查

应对数据集内或数据集间的研究变量或有效记

录进行唯一性评价，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。

5.1.4 一致性核查

应对不同数据集间的一致性进行评价。可核查现场调查与长期随访监测数据集间的个体唯一性标识以及数据标准是否一致。

5.1.5 准确性核查

应对数据内容的准确性进行评价，及时发现并纠正可识别的异常值或错误。

5.1.6 逻辑性核查

应对数据集内或数据集间的数据逻辑性进行评价，及时识别并纠正冲突值。

5.2 问题处置

5.2.1 补遗

存在缺失、异常值和错误的数据应经工作人员核实，并根据实际情况再次收集或重新测量这部分信息。

5.2.2 订正

5.2.2.1 不规范数据应依据统一的数据集标准进行订正。

5.2.2.2 对于异常、错误或逻辑冲突的数据，应经工作人员核实或再次收集该部分信息后在数据库中订正。

5.2.3 去重

重复数据应经工作人员核实，并选择性删除其中一条记录。

5.2.4 标准化及数据整合

对于多来源数据或不规范数据，应首先进行数据标准化及数据整合，宜参考本标准的第4和第6部分。

5.2.5 保留

若存在缺失数据无法填补或重复数据无法核实等暂时不可修改的问题时，应当记录并保留所有问题数据，在再次调查或随访时进行数据收集和确认，分批次处理上述问题。对于一些特殊问题，宜在条件许可或具备问题处理能力时开展专项调查，从根本上解决问题。

5.3 统计学监测

5.3.1 监测内容

为及时了解数据质量及数据库动态，应在数据清理的过程中定期进行统计学监测。统计学监测主要通过绘制数据获取进度图/表、数据分布图/表以及综合运用统计学分析方法对收集的队列数据进行分析和比较，识别可能存在的问题，及时向现场工作人员反馈，以提高队列研究的数据质量。监测内容包括数据获取进度、样本量及数据质量等情况。

5.3.2 监测方法

5.3.2.1 数据获取进度图/表

现场调查和长期随访监测过程中，应根据实际情况设定数据上报时限，并绘制数据获取进度图/表，及时记录每次数据的获取日期及样本量，便于掌握队列研究的工作进度和数据获取动态。

5.3.2.2 数据质控图

数据清理过程中，宜绘制适宜的质控图，客观地评价数据分布情况，实现统计异常的可视化。可选择散点图、直方图、折线图等方式，从以下五个方面展示数据分布特性，便于监测数据质量。

a）数据中心值的集中位置；

b）数据分布对称与否；

c）数据是否遵循特定分布规律；

d）数据分布中的峰形及峰值；

e）数据中的离群值。

5.3.2.3 逻辑检查

数据清理过程中，可通过计算均值或中位数、标准差、构成比、变异系数等统计学指标反映数据的分布情况和离散程度，识别数据中的异常值；通过缺失值和重复数据分析，了解数据集中缺失和重复数据的分布特征。每次数据清理完成后，应建立核查问题汇总表，可从规范性、完整性、唯一性、一致性、准确性和逻辑性等方面记录数据集中存在的相应问题，及时向现场工作人员反馈，采取相应的问题处置方法。

5.3.3 指标选择

5.3.3.1 准时率

依据数据获取进度图/表，可计算每次数据提交的准时率。当数据提交的准时率较低时，应及时联系现场工作人员，跟进现场调查或随访监测的进展，了解数据获取过程中存在的问题，给出相应的解决方法，必要时宜适当调整数据获取方案，保证数据可以准时提交。

准时率，即在规定时限内提交的数据集数目占所有提交数据集的百分比。

5.3.3.2 应答率及获取率

依据数据获取进度图/表，可在调查过程中及时监测研究对象的应答率或数据获取率，以掌握队列数据的样本变化情况和调查质量情况。

应答率是指所有被抽中的合格研究对象中有效参与调查研究的对象所占的比例。获取率是指该项调查中实际获取样本量占计划获取样本量的比例。

5.3.3.3 失访率

研究应长期动态地追踪研究对象的失访情况，以准确掌握队列人群的变化情况。

大型队列研究中，失访是指户口已迁出调查区域，且经查找仍无法得知去向，或虽有明确下落，但无法进行长期随访监测（如户口搬迁到外地等）的研究对象。当年研究人群的失访率，即为当年报告确认失访人数占同年随访人数的比例。

5.3.4 监测频度

研究人员应依据调查方案和实际情况合理设定监测频度。

6 数据整合与开发

6.1 基本内容

大型队列研究的数据整合，就是通过个体唯一性标识将大量的结构化数据和非结构化数据整合到标准数据库的过程。

6.1.1 结构化数据的整合

大型队列研究中常见的结构化数据形式包括现场调查数据和长期随访数据，通常具有较高的质量。数据整合过程需注意研究对象敏感信息和一般信息的区分，并与数据库中已有数据集建立连接。

6.1.2 非结构化数据的整合

大型队列研究的非结构化数据非常丰富，例如影像检查照片、录音文件等，推荐的整合形式有两种：

6.1.2.1 只保存原始文件和资料，并与已有的数据集建立连接；

6.1.2.2 在保存原始文件和资料的基础上，提取关键信息以结构化数据的形式整合到数据库。

6.2 基本过程

对大型队列研究的数据整合建议进行分步骤、分阶段管理，来应对项目实施的不同时期、不同阶段的数据管理需求。推荐将数据转化和整合过程分成四个阶段。

6.2.1 实时数据环境

实时数据环境存储和管理队列研究不同环节获取的实时、动态的原始信息，主要是用于项目现场运行和日常管理，数据需要进一步处理或清理才能用于研究。

6.2.2 数据开发环境

数据开发环境是在实时数据环境的基础上进一步整合其他离线数据和非结构化数据文件，该环境主要用于数据清理和整合。

6.2.3 数据分析环境

在数据管理人员、研究者和IT人员协作下，对相关数据进行处理形成统一的新变量用于后续的分析研究，例如计算量表得分、确定疾病诊断、衍生暴露综合变量等。在该阶段的环境中必须注意研究对象的隐私保护，去除研究对象的个人敏感信息。

6.2.4 数据分析固定环境

定期将数据分析环境形成固定版本，供研究团队内部和/或外部人员使用。数据分析固定环境须采用合适的数据格式、并配备必要的说明文档和数据使用协议，尤其强调研究对象的隐私保护和数据安全。

7 数据处理记录与报告

大型人群队列研究的数据处理可分为数据标准化、数据清理及质控和数据整合三个阶段，每个阶段任何涉及数据库的维护、更新、验证全历程的操作，都应详细记录数据处理过程、依据和结果。如有条件，宜对数据全过程做留痕处理，以便事后进行审计追责。数据处理结束后，应对数据处理工作和结果进行报告与评价。

7.1 计划

数据收集后，应立即备份原始数据，登记数据提交日期、文件名称及类型、样本量等基本信息，合理制定数据处理计划。数据处理计划宜从数据标准化、数据清理及质控和数据整合三个方面规定数据处理的目的、基本原则和具体流程。

7.2 执行记录

为确保数据处理工作的质量和可重复性，应严格执行数据处理计划，记录各阶段的执行时间、操作步骤、执行结果和其他关键信息。研究设计阶段，应记录数据库的建设框架和数据标准化过程。数据清理及质控阶段，应依据数据清理方案，记录数据清理及质控步骤，及时备份原始数据和必要文件。进行数据整合和开发时，应准备不同数据版本的更新日志和发布说明。

如因特定原因无法按原计划执行，经研究人员商议后，可合理修改数据处理计划，并按照新的计划开展后续工作。

7.3 报告和存档

数据处理工作完成后，应进行工作总结，报告数据处理各阶段的结果、可能存在的问题和相应解决办法，并且将原始数据、处理后数据以及必要文件进行归纳存档。

参 考 文 献

[1] WS/T 303－2009[S]卫生信息数据元标准化规则.

[2] WS/T 306－2009[S]卫生信息数据集分类与编码规则.

（收稿日期：2018-12-13）

（本文编辑：李银鸽）