

ICS **. ***, **
C**



团体标准

T/CACM ****-20**

中医真实世界研究数据采集技术规范

Standard for Data Acquisition in Real World Study of Chinese Medicine
(文件类型：送审稿)

20**-**-**发布

20**-**-**实施

中华中医药学会发布

目 次

(必备要素)

1 范围.....	1
2 规范性引用文件.....	1
3 术语和定义.....	1
4 选择数据源.....	1
5 评估数据.....	2
6 构建数据仓库.....	2
7 数据采集方法.....	3
8 数据采集要点.....	3
9 数据的提交.....	4
10 数据描述.....	4
附录 A (规范性附录) 研究文件归档与保存.....	5
参 考 文 献.....	6

前 言

本指南按照GB/T1.1—2020《标准化工作导则 第1部分:标准化文件的结构和起草》规定的规则起草。

本指南由中国中医科学院提出。

本指南由中华中医药学会归口。

本指南起草单位:中国中医科学院、中国中医科学院中医药数据中心、中国中医科学院中医药信息研究所、中国中医科学院中医临床基础医学研究所、安徽中医药大学第一附属医院、陕西中医药大学附属医院。

本指南主要起草人:谢琪、王斌、文天才、张磊、郭玉峰、刘保延、史华新、周洪伟、曹馨宇、忻凌、杜晓泉、张妮楠、张雯、张泽丹。

中医真实世界研究数据采集技术规范

1 范围

本规范规定了中医真实世界研究中数据源的选择、数据评估、数据仓库构建、数据提交、数据描述等数据采集的一般原则和方法。

本规范适用于临床研究机构从事中医真实世界数据采集人员。

2 规范性引用文件

下列文件对于本规范的应用是必不可少的。凡是注明日期的引用文件，仅所注明日期的版本适用于本规范。凡是不注明日期的引用文件，其最新版本（包括所有的修改版本）适用于本规范。

GB/T 15237.1—2000 术语工作 词汇 第1部分：理论与应用

WS/T 445—2014 电子病历基本数据集编制规范

T/CACM 1051-2017 中医真实世界研究技术规范通则

临床试验数据管理工作技术指南（国家食品药品监督管理总局2016年第112号通告）

临床试验的电子数据采集技术指导原则（国家食品药品监督管理总局2016年第114号通告）

3 术语和定义

下列术语和定义适用于本规范。

3.1

真实世界数据 (Real-world Data, RWD)

真实世界数据是指来源于日常所收集的各种与患者健康状况和/或诊疗及预防有关的数据。

3.2

真实世界证据 (Real-world Evidence, RWE)

真实世界证据是指通过对适用的真实世界数据进行恰当和充分的分析所获得的关于药物的使用情况和潜在获益-风险的临床证据，包括回顾性或前瞻性观察性研究或者实用临床试验等干预性研究获得的证据。

3.3

数据仓库

是面向主题的、集成的、相对稳定的、反映历史变化的数据集合，集成各种类型的应用系统，从发展的和历史的角来组织和存储数据，支持信息化和分析化处理。

3.4

电子数据采集 (Electronic Data Capture, EDC)

是一种基于计算机网络的用于临床试验数据采集的技术，通过软件、硬件、标准操作程序和人员配置的有机结合，以电子化的形式直接采集和传递临床数据。

4 选择数据源

除了研究者设计数据库用于收集研究数据之外，真实世界研究的数据还可来源于电子健康记录 (electronic health records, EHR)、社会保障或医疗保险数据库、疾病登记数据库、家庭监护设备数据、

移动健康设备的监护数据以及健康相关的纸质记录等。中医真实世界数据的主要来源于以电子病历（Electronic Medical Records, EMR）为核心的医疗卫生服务记录，以门诊病历或者住院病历的形式存储于医院信息系统的数据库中，主要包括中医的四诊信息、证候、病因病机、治则治法等，还包括一般人口学信息、诊断记录、处方、医嘱、检查检验报告以及移动健康设备的监护数据等。

5 评估数据

5.1 数据可行性评估

针对预设的临床问题，评估真实世界数据是否与所关注的临床问题密切相关。数据源的评估包括但不限于以下要素：

- a) 是否包含与临床结局相关的重要变量和信息，如药物暴露、患者人口学和临床特征、协变量、随访时间、结局变量等；
- b) 临床结局定义是否准确，相应的临床意义是否明确；
- c) 纳入的患者对于研究的目标人群是否具有代表性。

5.2 数据可靠性评估

真实世界数据的可靠性主要从数据的完整性、准确性、可溯源性和质量保证方面进行评价。

- a) 完整性：真实世界数据无法避免数据缺失问题，包括变量的缺失和变量值的缺失。应当尽可能保证研究方案确定的基线变量、关键变量的完整，尤其涉及研究的关键变量时，例如影响研究结局的诸多重要预后协变量缺失或变量值缺失，会加大研究结论的不确定性。
- b) 准确性：数据的准确性极为重要，通常需要参照较权威的数据来源进行识别或验证。数据元素和转化数据的算法均应保证其正确。数据的准确性还反映在数据的一致性和合理性上，一致性包括数据库内部的相关数据标准、格式和计算方法等必须一致；合理性包括变量数值的唯一性、合理的区间和分布、相关变量的预期依从关系以及时变型变量是否按预期改变等。
- c) 可溯源性：数据的来源、收集与治理的全过程应透明、清晰，并具有可溯源性，尤其是关键的暴露、协变量以及结局变量等应能追溯到源数据。数据的透明性还包括数据的可及性、数据库之间的信息共享和对患者隐私的保护方法的透明。
- d) 质量保证：数据采集应遵循统一的采集方法和流程采集数据，包括建立与收集真实世界数据有关的研究计划、协议和分析计划的时间安排、遵守采集关键数据点的共同时间框架；使用了共同定义数据标准；用于数据元素采集的技术方法是否充分，包括各种来源数据的集成、药物使用和实验室检查数据的记录、随访记录、与保险数据的链接以及数据安全等。

6 构建数据仓库

6.1 系统分析，确定主题

明确建立数据仓库所要解决的问题，分析数据来源，确定各个主题下的查询和统计分析要求。

6.2 选择满足数据仓库系统要求的软件平台

选择合适的软件平台，包括数据库、建模工具、分析工具等。需要考虑的因素包括：系统对数据量、响应时间、分析功能的要求等。

6.3 建立数据仓库的逻辑模型

6.3.1 确定建立数据仓库逻辑模型的基本方法。

6.3.2 基于主题视图，把主题视图中的数据定义转到逻辑数据模型中。

6.3.3 识别主题之间的关系，分解多对多的关系。

6.3.4 用范式理论检验逻辑数据模型。

6.3.5 由用户审核逻辑数据模型。

6.4 数据仓库设计的文档要求

数据仓库设计文档至少包括 3 个部分，即：引言、外部设计和结构设计。引言包括：编写说明、背景、修订审批记录；外部设计包括：标识符和状态、使用该数据仓库的程序、数据仓库中表（Table）的命名约定、数据仓库表中字段（表的列 Column）命名约定、数据仓库中创建的视图（View）命名约定、数据库中的数据仓库链接（DATABASE LINK）的命名约定、输入参数及输出参数命名约定、变量命名约定；结构设计：概念结构设计、逻辑结构设计、表目录、表结构、物理结构设计、设计数据的存取路径、设计数据的存放位置、完整性设计、安全保密设计、用户身份的标识、存取控制。

6.5 数据仓库测试

测试数据仓库之前制定测试计划书，测试系统功能是否与前期设计及说明书一致。测试内容包括：浏览及录入页面设计，各个访视顺序、访视中的录入表格顺序及每个数据点的顺序；不同用户浏览权限的准确性等等。

6.6 数据清洗转换和加载

业务系统中的数据在加载到数据仓库之前，必须进行数据的清洗和转换，保证数据仓库中数据的一致性。加载方案必须能够支持访问不同的数据库和文件系统；数据的清洗、转换和传输必须满足时间要求，能够在规定的时间范围内完成；支持各种转换方法，各种转换方法可以构成一个 workflow；支持增量加载，只把自上一次加载以来变化的数据加载到数据仓库。

7 数据采集方法

7.1 回顾性数据采集

回顾性数据的采集，首先要提交回顾性研究方案，并通过伦理审查；其次是提交数据使用申请；然后才能通过专用的检索、查询模块，根据纳入、排除标准，设置约束条件，利用导出功能，批量导出数据。

7.2 前瞻性数据采集

和常规前瞻性研究类似，数据在收集之前已有明确的研究目的，制定好了数据采集模板或 CRF 表，数据采集一般通过电子数据采集系统来完成，因此数据质量较高。

8 数据采集要点

8.1 明确数据采集范围

- a) 时间范围，明确数据集的标志性时间点，如门急诊患者以就诊日期为约束，住院患者以出院日期为准，格式按年上报 yyyy，按月上报 yyyyymm，按日上报 yyyyymmdd；
- b) 内容范围，明确数据集的变量范围，如患者门诊，住院，诊断，治疗，检查，检验，随访，中医康复，专科特色疗法，病案首页数据，特殊疾病数据，不良事件等。

8.2 多源异构数据的汇交问题

对于不同来源数据库的汇交。首先，建立统一的数据标准，将不同来源的数据进行标准化处理。其次，每一条记录都应有唯一标识码。推荐使用医保手册号或者身份证号作为各数据的唯一标识。

8.3 数据的隐私问题

数据源获取方法及数据的检索策略必须公开、透明，并必须通过伦理审查委员会的审核，对于没有足够的匿名化处理方法达到隐私保护要求的数据源，应舍弃。

9 数据的提交

电子数据能够定期从各研究中心传输给发起者，如使用文件传输协议（FTP）完成数据文件的提交，也可以通过注册登记平台进行病例数据的录入。通过纸质数据采集表获取的数据，可以将表格寄送给发起者的集中电子数据处理机构，由中心统一录入。

10 数据描述

数据特征描述：数据来源、数据量。

数据库描述：数据表的定义规则、变量定义规则、隐私保护要求、共享要求。

数据文件描述：文件格式、存储方式、备份策略、保存时限、数据库说明文档。

数据用户描述：临床数据采集者、质量控制人员、数据分析人员。

CACMVA 临床数据

附录 A
(规范性附录)
研究文件归档与保存

序号	应保存的档案	研究者	数据所有权单位
1	临床研究单位及主要研究者基本情况	保存	保存
2	数据使用协议	保存 (原件)	保存 (原件)
3	数据采集表 (样表)	保存	保存
4	数据描述报告	保存	保存
5	统计评估报告	保存	保存
6	数据仓库及说明文档等	保存	保存

CACM 公共数据库

参考文献

[1] Junhua Zhang, Boli Zhang. Clinical research of traditional Chinese medicine in big data era [J]. *Frontiers of Medicine*, 2014, 8(3): 321-327.

[2] Food and Drug Administration. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff[EB/OL]. [2017-8-31].

<https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm513027.pdf>.

[3] 刘保延. 真实世界的中医临床科研范式. *中医杂志*[J]. 2013, 54(6): 451-455.

[4] 卜擎燕, 熊宁宁, 邹建东等. 从临床研究数据管理角度设计病例报告表[J]. *中国新药杂志*, 2007, 16(5): 339-343.

[5] 孙鑫, 谭婧, 唐立等. 重新认识真实世界研究[J]. *中国循证医学杂志*, 2017, 2(17): 126-130.

[6] 谢琪, 江丽杰, 刘保延, 等. 开展真实世界中中医药效果比较研究的关键问题及对策的探讨[J]. *世界中医药*, 2014(1):28-31.

[7] Sherman, R. E., et al., Real-World Evidence - What Is It and What Can It Tell Us? *New England Journal of Medicine*, 31: 2016, 375(23).

[8] 王斌, 谢琪, 史华新, 等. 真实世界临床科研中电子病历数据的存储[J]. *中医杂志*, 2013, 54(20): 1736-1739.

[9] GavinPowell. 数据库设计入门经典[M]. 清华大学出版社, 2007.