.*.**
C**

团体标准

T/CACM ****—20**

中医药真实世界研究技术规范 数据库构建和数据预处理

Technical Specifications for Real-World Studies of Traditional Chinese Medicine

Database Construction and Data Preprocessing

（文件类型：公示稿）

（完成日期：2020 年 6 月）

20**-**-**发布　　20**-**-**实施

中华中医药学会发布

目 次

前　言

《中医药真实世界研究技术系列规范》包括如下标准：

——T/CACM ***.1 中医药真实世界研究技术规范 数据库构建和数据预处理；

——T/CACM ***.2 中医药真实世界研究技术规范 统计分析计划制定；

——T/CACM ***.3 中医药真实世界研究技术规范 证据质量评价与报告；

——T/CACM ***.4 中医药真实世界研究技术规范 伦理审查；

——T/CACM ***.5 中医药真实世界研究技术规范 基于证据的中药有效性及安全性评价；

本文件是该系列规范的第 1 个规范。

本文件按照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。

本文件由北京大学、中国中医科学院西苑医院、中药临床疗效和安全性评价国家工程实验室、中药临床研究与评价重点实验室提出。

本文件由中华中医药学会归口。

本文件起草单位：北京大学、中国中医科学院西苑医院、中药临床疗效和安全性评价国家工程实验室、中药临床研究与评价重点实验室、北京中医药大学、广州中医药大学第一附属医院、温州医科大学、广西医科大学、中国中医科学院中医临床基础医学研究所、北京积水潭医院、汕头大学、北京大学肿瘤医院

本文件主要起草人：陈大方、高蕊、闫泽玉、訾明杰、孙明月、陆芳、费宇彤、杨忠奇

本文件起草人：毛广运、余红平、车前子、武轶群、段芳芳、郭貔、刘志科、马逸杰、周泽宸

引　言

真实世界数据来源于真实的诊疗过程而非特定研究背景。真实世界研究是指针对预设的临床问题，在真实世界环境下收集与研究对象健康有关的数据或基于这些数据衍生的汇总数据，通过分析，获得药物的使用情况及潜在获益-风险的临床证据的研究过程。在中医药方面，通过利用真实世界数据开展中医药真实世界研究，为药物有效性和安全性评价提供真实世界证据。2020 年 1 月国家药品监督管理局发布《真实世界证据支持药物研发与审评的指导原则（试行）》，明确真实世界证据可用于支持药物研发与审评。这也对真实世界研究质量提出了要求。但是，由于中医药真实世界数据具有增速快、来源广、容量大和复杂性的特点，在数据规范化程度、多源信息整合和安全隐私保护方面存在不足，阻碍了中医药领域真实世界数据转换为真实世界证据。因此，亟需对数据采集、提取、合并、预处理等方面建立技术规范，提高数据质量，更好规范中医药真实世界研究的开展。

本规范基于中医药真实世界数据，结合《真实世界证据支持药物研发与审评的指导原则（试行）》和《用于产生真实世界证据的真实世界数据指导原则（试行）》等技术文件对真实世界数据的要求，建立数据库构建和数据预处理的技术规范，以保证真实世界研究结果的真实性和可靠性，提高中医药真实世界研究质量。

中医药真实世界研究技术规范 数据库构建和数据预处理

1 范围

本文件规定了中医药真实世界研究中基于既有真实世界数据构建用于分析研究的数据库，以及数据预处理的原则和规范。

本文件适用于中药临床评价及临床科研、药物审批机构、中药企业、中医院、中西医结合医院、民族医院、综合医院、高校科研人员使用。

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）适用于本文件。

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

T/CACM 022—2017 中医真实世界研究技术规范通则

国家药监局 2021 年第 27 号 用于产生真实世界证据的真实世界数据指导原则（试行）

国家药监局2020年第1号通告 真实世界证据支持药物研发和审评的指导原则(试行）

国家药监局 2020 年第 16 号通告 药物临床试验数据递交指导原则（试行）

国家药监局 2020 年第 57 号通告 药物临床试验质量管理规范

国家药监局 2016 年第 112 号通告 临床试验数据管理工作技术指南

国家药监局 2016 年第 114 号通告 临床试验的电子数据采集技术指导原则

吴阶平医学基金会，中国胸部肿瘤研究协作组 2018 年 《真实世界研究指南（2018 年版）》

3 术语及定义

下列术语和定义适用于本《规范》。

3.1

真实世界研究 Real-World Research/Study，RWR/RWS

针对临床研究问题，在真实世界环境下收集与研究对象健康状况和/或诊疗及保健有关的数据（真实世界数据）或基于这些数据衍生的汇总数据，通过分析，获得药物的使用价值及潜在获益-风险的临床证据（真实世界证据）的研究过程。

[来源：国家药监局 2021 年第 27 号 用于产生真实世界证据的真实世界数据指导原则（试行）]

3.2

真实世界数据 Real-World Data，RWD

来源于日常所收集的各种与患者健康状况和/或诊疗及保健有关的数据。并非所有的真实世界数据经分析后就能成为真实世界证据，只有满足适用性的真实世界数据才有可能产生真实世界证据。

[来源：国家药监局 2021 年第 27 号　用于产生真实世界证据的真实世界数据指导原则（试行）]

3.3

既有数据 Existing Data

真实世界中已经存在的数据，常见来源包括医院信息系统数据、医保支付数据、疾病登记数据、公共卫生监测数据（如药品安全性监测、死亡信息登记、院外健康监测）、自然人群队列数据等。

[来源：国家药监局 2020 年第 27 号通告 用于产生真实世界证据的真实世界数据指导原则（试行），有修改]

3.4

原始数据库 Original Database

研究设计，来源于真实世界既有数据，能够直接用于后续数据预处理和分析的数据库。依据研究目的和纳入排除标准，提取既有数据库相关变量和记录形成的数据库

构建数据库即指构建原始数据库。

[来源：国家药监局 2020 年第 16 号通告 药物临床试验数据递交指导原则（试行），有修改]

3.5

分析数据库 Analysis Database

对原始数据库进行数据预处理之后，形成的可直接用于后续统计分析的数据库。

[来源：国家药监局 2020 年第 16 号通告 药物临床试验数据递交指导原则（试行），有修改]

3.6

内部标识符 Internal Identifier

又称为内部标识变量、主键。是指用于识别单个数据集内每一条记录的变量。如患者住院号、患者序列号等。

3.7

外部标识符 External Identifier

又称为外部标识变量、外键。是指每个数据集中用于链接不同数据集的标识变量。

3.8

数据预处理 Data Preprocessing

数据清理、数据转换等数据预处理的过程。通过数据预处理，以确保数据规范、提高数据质量，形成适用于下一步统计分析或数据挖掘及可视化的数据集。

3.9

数据质量 Data Quality

包括但不限于以下方面：完整性、准确性和可溯源性。完整性用于评价是否包含相关研究变量和研究人群，以及变量的缺失情况。准确性用于评价数据与其描述的客观实体的特征是否一致。可溯源性用于评价数据使用全过程是否存在相关记录，以保证分析可重复性。

4 原始数据库构建与评价

4.1 一般构建流程

4.1.1　构建研究方案，确定研究目的、纳入排除标准等;

4.1.2　依据研究目的，建立相关变量清单;

4.1.3　选择单个（多个）既有数据来源;

4.1.4　从选择的既有数据来源中，提取研究相关变量，并依据研究纳入排除标准，提取研究对象的相关记录，形成原始数据库。为确保数据的完整性和真实性，建议尽可能利用多个既有数据库构建原始数据库。

4.2 既有数据选择原则

4.2.1 数据可获得

既有数据可以开放或者共享。

4.2.2 数据可靠性

数据记录及时、真实、准确。

若既有数据存在数据质量控制标准，应满足研究要求。若既有数据不存在数据质量控制标准，建议对数据进行溯源，了解数据实际采集情况，判断是否满足研究要求。如中医的病、证、症、征及体质等中医诊断标准是否明确，诊断方式是否可靠。不同临床医师的诊断水平是否一致，不同诊断设备的诊断结果是否统一，以及用语的规范性。

4.2.3 数据完整性

包含研究所需的关键变量，即反映干预/暴露相关变量、结局变量和协变量。

诊疗数据应包含生成时间，实验室检查结果应附正常参考值，涉及的干预措施应有相应鉴别信息，诸如中成药记录应附生产厂家及批号。

4.2.4 伦理学规范

数据使用符合伦理学规范。

4.3 数据提取原则

4.3.1 提取完整性

包括变量完整性和记录完整性。

提取变量至少包含内部标识变量、反映干预/暴露相关变量、结局变量和协变量。如果为多个既有数据来源，还应包含外部标识变量。各变量信息必须包括变量名称、含义、数据类型和数据采集方式。对于中医药研究还应至少包含：

1）所研究药物的剂型和煎煮方式。例如中药饮片自煎或代煎、院内制剂的制备工艺等；

2）所研究药物的组成及剂量，例如固定处方/协定方、经典方、院内制剂、名老中医经验方等，应包含药物组成及加减药物的范畴、各味中药的具体剂量或增减剂量范畴。

应评估和确立提取的字段，制定相应的核查规则，以判断提取的记录是否完整。

4.3.2 提取规范性

在数据提取前，形成数据提取方案。数据提取方案至少包括提取变量及变量信息、数据来源、提取方法和操作人员。

数据提取应有符合资质、具有相应设备的专业人员或机构进行。

4.3.3 提取准确性

数据提取应保证准确性，即提取到的数据与既有数据库中的数据一致，以保证数据的真实性。

4.3.4 可溯源性

保留数据提取方案，记录数据提取来源和实际提取过程。记录临床数据的原始值，需附数据源、产生时间及操作者，对数据的任何修改，需附日期和时间、修改原因、操作者。

4.3.5 伦理要求

提取方案符合伦理规范，敏感信息应匿名化处理。

4.4 多源数据合并

多源数据包括一个机构内部的多源数据，不同医疗机构的数据等，研究者应尽量提取到患者在研究周期内的全部诊疗数据。

由多个既有数据分别提取的原始数据库，应包含唯一的外部标识变量。

如果不同来源的既有数据库存在数据重复/矛盾时，建议通过评估既有数据的采集方式和质量控制标准与研究目的是否相符，建立重复/矛盾数据优先级。如电子健康档案的个人信息数据相对于来自移动设备端可能更准确。

在合并多源数据库前，建立统一的数据规范和操作标准，对数据进行规范化处理后再予以合并。数据合并规范应至少包括以下内容：

1）规定最终数据库中变量名称、定义、度量单位。如同来源数据库中身高数据分别采用厘米（cm）和米（m），应转换为相同单位后再予以合并；

2）明确数据采集方法和测量方法，以及不同方法获得的数据是否可以合并。如实验室检验数据可由不同设备、机构检测获得，应判断是否可以合并，或者是否可以经转换后合并；

3）明确数据提取和合并的方法和流程。

4.5 隐私保护

研究方案和数据采集/提取方案需经过伦理委员会审查。数据采集/提取、数据合并等阶段均需设计患者隐私保护方案，保护患者个人隐私免遭泄露。

应明确原始数据库共享范围、人员、形式和期限。

具体内容参考本系列规范《真实世界数据用于真实世界研究的伦理审查规范》。

4.6 质量控制

质量控制应贯穿数据处理的每一个环节，涉及数据提取、安全处理、清洗、结构化等环节，确保数据的真实性、准确性和完整性。应依据研究设计，依托于相应的数据质量评价，制定质量控制方案，考虑的内容包括但不限于：

1） 是否建立与真实世界数据有关的研究计划、方案和统计分析计划；

2） 确保既有数据的准确性和真实性：既有数据质量控制标准应满足分析要求；来源于门诊的疾病描述、诊断及其用药信息需要有相关证据链佐证；任何修改应有负责人的确认，确保留下完整的稽查轨迹；

3） 数据提取、多源数据合并等方面是否有相应的标准操作规程；数据提取是否有明确流程和合格人员等；

4） 是否依照预先规定的流程进行；

5） 是否满足可溯源性要求，保存数据处理的所有记录。

4.7 评价

4.7.1 数据提取评价

4.7.1.1 准确性评价

评估提取到的数据与既有数据是否一致，可采用人机结合的方式，利用专门的核查工具或人工核对记录数，进行数据时间校验，或者随机抽取一定比例的数据与原始数据进行人工核对。如出现不一致的情况，需分析出现原因并予以纠正。

4.7.1.2 规范性评价

评估数据提取和多源数据合并是否符合预先规定的数据规范和操作标准。

4.7.1.3 完整性评价

评估是否已提取关键变量、变量信息和所有记录。可采用核对记录数、数据时间校验等方式。

涉及多源数据合并，需要评估外部标识变量在各个数据库中的完整性和链接比例，预估在此链接比例下经数据预处理后的样本量能否满足研究要求。

4.7.2 原始数据库的数据质量评价

4.7.2.1 完整性评价

数据库人群数量应该满足研究样本量要求。人群特征应能满足研究所需，如年龄分布、性别构成、合并症种类、辅助治疗种类等；

应包含研究关键变量，即反映干预/暴露相关变量、结局有关变量和协变量。主要研究变量缺失比例应在允许的范围内；

个案记录的观察时间应该满足研究所需，即包含治疗期和随访期。

4.6.2.2 准确性评价

评价数据与其描述的客观实体的特征是否一致，即数据是否准确、真实。可依据既有数据的质量控制标准或者数据实际采集情况进行评估，判断其是否符合研究要求。

4.6.2.3 可溯源性评价

研究数据均能够追溯至源数据，即要求保留数据处理方案和数据处理具体记录。包括记录选择的既有数据库，数据提取方案、数据合并、数据筛选的方案和过程。

5 数据预处理与评价

5.1 总原则

数据预处理应在真实性和可溯源性的前提下，选择合理的数据预处理技术，处理缺失数据、重复数据和异常数据等，对数据进行规范化和结构化处理，提高数据质量，满足数据的一致性、唯一性、完整性和准确性的要求。

5.2 数据清理

5.2.1 重复值

由于不同数据来源引起的变量重复时，建议依据既有数据的数据采集方式和质量控制标准，建立重复数据优先级，删除重复变量。

存在完全重复的记录，应予以删除。

由于内部标识编码重复导致数据重复时，应经人工核实后予以纠正。

5.2.2 异常数据

常见的异常数据包括逻辑错误数据和离群值等。

逻辑错误常见于变量之间不符合逻辑关系，例如出院时间早于入院时间、实验室定性判断结果与方案中定义的判断标准不一致等。应进一步核实后才能修订数据，数据的修订应保留记录。

离群值依据产生原因可分为两类：人为离群值和自然离群值。自然离群值建议采用稳健统计方法进行分析。人为离群值需经核实后进行纠正。

5.2.3 缺失值

依据缺失值产生的原因，缺失值可分为完全随机缺失，随机缺失和完全非随机缺失。在对缺失值进行处理时，需判断变量的缺失是否能够采用可靠的统计学方法进行处理，再依据缺失原因选择合适的处理方法，并采用敏感性分析等方法判断缺失处理方法对结果的影响。

完全随机缺失和随机缺失可采用删除法和插补法。此外，如果后续数据分析采用构建模型方法，随机缺失可以不用事先插补，模型估计时采用能够处理缺失的估计方法即可，如EM算法。

5.2.4 一致性

一致性包括但不限于：记录格式统一性、编码统一性。如对于时间变量，均采用“年/月/日”的记录格式。

中西医诊断编码、四诊信息、中药品名称、手术名称、药品不良反应编码等规范应参考标准。标准权威性依次为国际/国家/地方标准、行业标准、字/词典、教材。如国际疾病分类编码 （international classification of diseases，ICD）、《GB/T 20348-2006 中医基础理论术语》、《GB/T 16751.1-1997 中医临床诊疗术语 疾病部分》、《GB/T 16751.2-1997 中医临床诊疗术语 症候部分》、《GB/T 20348-2006 中医基础理论术语》、《中药饮片调剂规范》、《中华人民共和国药典》等。

5.3 数据转换

数据转换包括自然语言处理、衍生变量计算等。非结构化数据，如中医的病、证、症、征、体质、舌诊、脉诊等数据，可采用自然语言处理方法，如基于事件、变量字典，通过文本标注、分词等，进行事件、实体、变量等的识别和语义关联建立，也可通过人工标注等方法进行。在保证准确性和可溯源下，进行结构化处理。

5.4 分析数据库的数据质量评价

对原始数据库进行预处理后的数据库为分析数据库。对分析数据库的数据质量评价包括一致性、唯一性、完整性、准确性和可溯源性，如下：

——一致性：数据记录格式和编码规则统一；

——唯一性：无重复记录和重复变量；

——完整性：至少包含研究相关变量。研究变量的缺失情况应满足代表性和模型分析要求。如采用填补法处理缺失值，需评估缺失处理方法对结果进行因果推断的影响；

——准确性：变量值在合理区间范围内；

——可溯源性：应保留原始数据和数据清理过程记录。

参 考 文 献

[1] FDA. Developing a framework for regulatory use of real-world evidence; Public Workshop. https://www.gpo.gov/fdsys/pkg/FR-2017-07-31/pdf/2017-16021.pdf. Accessed 08 Sep 2017.

[2] Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Adv Ther. 2018;35(11):1763-1774. doi:10.1007/s12325-018-0805-y.

[3] Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. J Healthc Eng. 2018;2018:4302425. Published 2018 Apr 8. doi:10.1155/2018/4302425

[4]王雯,高培,吴晶,宣建伟,贺小宁,胡明,李洪超,窦丰满,于川,闫盈盈,孙鑫.构建基于既有健康医疗数据的研究型数据库技术规范[J].中国循证医学杂志,2019,19(07):763-770.

[5]彭晓霞,舒啸尘,谭婧,王丽,聂晓璐,王雯,温泽淮,孙鑫.基于真实世界数据评价治疗结局的观察性研究设计技术规范[J].中国循证医学杂志,2019,19(07):779-786.

[6]卞铮,许祥,余灿清,韩晓,俞敏,龚巍巍,吕筠,刘亚宁,郭彧,李立明.大型人群队列现场调查管理技术规范团体标准解读[J].中华流行病学杂志,2019(07):753-755.

[7]龚巍巍,俞敏,郭彧,王蒙,吕筠,余灿清,卞铮,王浩,谭云龙,裴培,李立明.大型人群队列终点事件长期随访技术规范团体标准解读[J].中华流行病学杂志,2019(07):756-758.

[8]余灿清,李立明.大型队列研究中的数据科学[J].中华流行病学杂志,2019(01):1-4.

[9]余灿清,刘亚宁,吕筠,卞铮,谭云龙,郭彧,汤海京,杨旭,李立明.大型人群队列研究数据管理团体标准解读[J].中华流行病学杂志,2019(01):17-19.

[10]胡文, 侯政昆, 刘凤斌,等. 关于大数据时代的中医药临床研究的思考[J]. 世界科学技术-中医药现代化, 2019.

[11] 张冬,张俊华,孙凤,田金徽,庞稳泰,杨丰文,金鑫瑶,刘智,郑文科.真实世界研究与中医药大数据[J].世界中医药,2019,14(12):3119-3122.

[12]陈大方,刘徽. 医学大数据挖掘方法与应用[M]. 北京大学医学出版社, 2020.

[13]高培,王杨,罗剑锋,任燕,胡明,唐少文,胡皓,孙鑫.基于真实世界数据评价治疗结局研究的统计分析技术规范[J].中国循证医学杂志,2019,19(07):787-793.