

真实世界数据库构建标准指南

目录

一、总则	- 2 -
(一) 背景与目的	- 2 -
(二) 适用范围	- 2 -
(三) 基本原则	- 3 -
1. 科学性	- 3 -
2. 真实性	- 3 -
3. 合规性与安全性	- 4 -
4. 高质量	- 4 -
(四) 组织管理与职责	- 4 -
二、真实世界数据库构建规范	- 5 -
(五) 数据采集范围	- 5 -
1. 数据源范围界定	- 5 -
2. 数据形态与多源异构特性	- 6 -
(六) 构建业务镜像库	- 7 -
1. 数据治理流程	- 7 -
2. 数据质控	- 8 -
(七) 构建标准资源库	- 9 -
1. 数据治理流程	- 9 -
2. 数据质控	- 10 -
(八) 构建观察数据库	- 10 -
1. 数据治理流程	- 10 -
2. 数据质控	- 11 -
(九) 构建科学变量库	- 11 -
1. 数据治理流程:	- 12 -
2. 数据质控	- 12 -
(十) 构建科学数据库	- 13 -
1. 数据治理流程:	- 13 -
2. 数据质控	- 14 -
三、安全与隐私保护	- 14 -
(十一) 隐私保护技术	- 15 -
(十二) 数据安全运维	- 15 -
四、术语与定义	- 16 -
(十三) 真实世界研究 (RWS)	- 16 -
(十四) 真实世界数据 (RWD)	- 16 -
(十五) 真实世界证据 (RWE)	- 17 -
(十六) 通用数据模型 (CDM)	- 17 -
(十七) JDBC (Java Database Connectivity)	- 17 -
(十八) 患者主索引 (Enterprise Master Patient Index, EMPI)	- 18 -

真实世界数据库构建标准指南

一、总则

本指南旨在确立一套标准化、可量化、可追溯的数据治理与分析体系，以应对多源异构医疗数据在科研转化过程中的复杂挑战，确保数据资产的规范化管理与高效利用。

（一）背景与目的

在当前的医疗大数据应用背景下，原始的真实世界数据（RWD）虽然体量巨大，但往往分散在医疗、医保、药监等不同的业务系统中，存在格式不一、标准缺失、颗粒度粗糙等共性问题。本指南的制定，通过建立规范化的操作流程与多维度的质控体系，重点解决数据采集进度失控、数据质量参差不齐、关键数据缺失与错乱，以及既定数据标准未履行等核心问题。

（二）适用范围

本指南适用于多源、异构、跨机构的医疗健康数据全流程管理，具体涵盖以下维度：

全覆盖的数据采集对象：本指南适用于全国范围内的“三医联动”数据治理体系，涵盖卫生健康部门各级医疗机构的临床诊疗数据、医疗保障部门医保结算与费用数据以及药品监管部门药品不良反应与追溯数据。数据对象包括但不限于医院信息系统（HIS）、实验室系统（LIS）、影像系统（PACS）及医保结算平台中的业务数据。

数据类型与技术形态：本指南适用于处理各种复杂形态的异构数据。既包括结构化的关系型数据库记录，也包括非结构

化临床文书。在技术治理架构上，本指南规定了从构建业务镜像库（BMDB）进行原始备份，到基于资源通用数据模型（RCDM）构建标准资源库，再到基于活动模型构建观察数据库，最终形成面向全域人群的科学数据队列仓库（SCDM）的全链路操作规范。

应用场景：本指南指导下的数据资产，旨在服务于广泛的医学与管理目标。适用于流行病学研究、临床试验设计辅助、真实世界研究（RWS）、精准医学分析、卫生技术评估以及智慧医疗和数字健康应用的开发，不限特定病种，旨在构建通用的科学数据基础设施。

（三）基本原则

1. 科学性

数据治理严格参考国际国内行业标准，包括 ICD-10、ICD-9-CM3、YPID、CNAS-AL09 及《WS364 卫生信息数据元值域代码》等。

依托科学通用数据模型（SCDM），构建包含 143,121 个元数据变量的庞大变量体系。数据组织方式采用“患者标识-记录时间-发生时间-变量 X”的纵向排列结构，直观表达科学数据。

在数据标准化归一过程中，机器归一的准确率和召回率均要求达到 90%以上。

2. 真实性

坚持业务镜像库原则，即镜像库数据模型与源业务信息系统完全相同。

建立临床文档浏览器，支持从清洗后的数据回溯浏览原始病历文档，确保任何数据点均有据可查。

所有的脱敏操作均不可逆，且采集过程需记录详细日志，防止数据被人为篡改或伪造。

3. 合规性与安全性

对数据须进行脱敏处理，保护个人隐私及数据对象利益。脱敏后数据不能进行反脱敏，仅在特定权限下通过物理隔离的对照机制进行必要的反向追溯。

采集过程应采用增量更新方式，只采集自上次采集以来发生变化的数据，最大程度减少对源业务系统的资源消耗和依赖。

4. 高质量

建立严格的质控规则，从规范性、完整性、准确性、一致性以及时效性这五类指标对数据质量进行评价。

数据治理平台需与数据生产、更新工作协同推进，构建原生数据优化闭环与数据清洗能力提升闭环双体系，实现数据质量的持续迭代与优化。

（四）组织管理与职责

为了保障数据治理体系的高效运转，明确各方在流程中的具体职责：

数据所有者：负责提供数据接入授权，界定数据的业务含义。医保局、卫健委及医院等医疗机构作为数据源头，需配合建立针对各细分业务域的质控规则库。

技术建设方：使用 JDBC/ODBC/RESTful API 建立连接，编

写并维护定期运行的采集脚本；负责构建业务镜像库、标准资源库及科学数据仓库；实施患者主索引（EMPI）的算法配置，基于身份证、姓名等信息实现跨机构患者身份的交叉匹配。维护数据治理平台，确保数据处理过程的完整性与一致性。

临床研究者：参与构建科学变量库，根据医学主题定义变量的概念与值域；对系统无法自动映射的观察项目（如特殊体征、非标手术名称）进行人工标注说明，并对机器归一结果进行核查，计算准确率。

数据分析师与运营人员：运用数据挖掘技术和医学信息规则引擎，建立多维度质控体系；针对系统生成的质量问题列表，排查确认原生数据问题，剖析问题发生原因，并提交问题复现路径；负责生成包含综合数据质量评分和业务主题质量评分的数据质量报告。

二、真实世界数据库构建规范

（五）数据采集范围

1. 数据源范围界定

针对真实世界研究的需求，应对以下三大类多源异构数据进行详细评估与接入规划：

医疗机构数据：侧重于临床诊疗过程与结果，是评价有效性的核心。采集的核心系统包括医院信息系统（HIS）、实验室信息系统（LIS）、医学影像存档与通讯系统（PACS）、电子病历系统（EMR）等，关键数据元应包含患者基本信息、门诊/住院记录、病案首页、医嘱明细、检验检查结果、手术操作记录、护理记录等。

医疗保障数据：侧重于医疗费用与资源消耗，是药物经济学评价的基础，具有极高的随访完整性，可用于追踪患者的长期生存状态及医疗资源消耗情况。采集的核心数据包括医保结算清单、费用明细、医保版疾病诊断编码、医保药品目录匹配数据、异地就诊结算数据等。

药品监管数据：侧重于产品全生命周期安全性与追溯。采集的核心数据包括药品不良反应监测数据、药物警戒信号、药品流通追溯数据、药械注册信息等。

其他数据：侧重于补充患者的社会人口学特征、生存结局及社会支持信息，是完善患者全生命周期数据链条的重要补充。采集的核心数据包括人口基础信息、户籍登记数据、死亡注销数据、婚姻登记数据、社会救助信息、殡葬服务数据、区域人口统计数据、环境监测数据等。此类数据可用于补充医保随访缺失的死亡结局信息、分析社会经济因素与环境因素对疾病发生发展及预后的影响，提升真实世界研究结论的科学性与全面性。

2. 数据形态与多源异构特性

数据采集对象具有多结构、多格式特征，需重点应对多源异构问题。结构化数据主要来自关系型数据库存储的业务表，包含患者基本信息、医嘱记录等；非结构化/半结构化数据包含病历文书、出院小结、病理报告、CT/MRI 影像等类型。数据域需实现全面覆盖，具体涵盖社会/经济/环境/地理、疾病、手术、药品、检验、术语（含影像）、中医症候七大核心域，保障数据维度的完整性与全面性。

（六）构建业务镜像库

为了从多源异构数据源采集数据，并最大程度地减少对数据源业务系统的依赖和影响，需要对不同数据源的不同系统、不同厂商的业务数据库中的数据镜像（复制）并且融合至同一个数据库，形成业务镜像库（BMDB）。业务镜像库实际上是对数据源多源异构业务数据库的备份，其数据模型与业务信息系统相同。

业务镜像库是多源异构数据源原始数据的备份，其数据模型与业务信息系统相同。业务镜像库的数据主要来源于全省医疗机构。

1. 数据治理流程

（1）建立连接

针对“三医”领域数据源系统厂商众多、技术架构各异现状，采用适配性强的连接技术以打通数据传输通道。项目组应根据各数据源系统的实际环境，灵活选用 JDBC 或 RESTful API 等标准技术来建立稳定的数据连接。此步骤不仅涉及物理网络的连通，还需解决异构数据库的兼容性问题，确保采集端能够准确识别并读取源系统的底层数据结构。

（2）编写采集脚本

数据采集不是一次性的动作，技术团队需深入分析数据源系统的数据库结构和具体需求，编写定制化的数据采集脚本并部署在统一的调度平台上定期自动运行，让系统能够定时从数据源系统中精准抓取所需数据，无需人工干预。

（3）增量更新机制

考虑到全省医疗数据体量巨大，为了大幅提高采集效率并减少网络带宽与计算资源的消耗，系统应强制采用增量更新的方式。增量更新的核心逻辑是只采集变化量，即系统通过比对时间戳，仅抓取自上次采集任务完成以来发生新增或状态变更的数据。

（4）患者主索引构建（EMPI）

建立全局唯一的患者主索引（EMPI ID），作为跨系统数据关联的主键。系统采用以身份证（或社保卡）为核心标识符、辅以姓名等信息的多维度匹配算法，利用该 EMPI ID 整合医保、卫健、药监等系统的数据，最终构建完整的患者全周期诊疗数据链条。

（5）数据脱敏

为了保护个人隐私，保护数据代表的对象利益不受侵害，需要对数据进行脱敏处理，所有的脱敏均不可逆，数据脱敏后不能进行反脱敏。对于患者信息，在个人隐私信息脱敏后，可以通过相关方法进行反向追溯。

2. 数据质控

业务镜像库质控是针对业务数据库向镜像库备份同步过程的专项质量控制，核心是验证备份数据的完整性与一致性，检查是否存在数据缺失、遗漏、损坏或同步延迟等问题，确保镜像库是业务生产库的精确可用副本，为后续科研分析、数据查询等非生产场景提供可靠基础。

开展业务镜像库质控，需建立定期校验机制，每日比对两库的表结构、记录总数及核心关键字段值，定期执行全量数据

哈希校验；同时实时监控增量同步链路状态，设置数据差异阈值告警；发现异常时追溯备份日志定位问题，完成数据补全修复后重新验证，并定期复盘优化备份与质控规则。

（七）构建标准资源库

依托镜像数据库，遵照资源通用数据模型（RCDM），对多源异构的业务数据进行组织和存储，形成标准的、模型统一的健康数据资源，构建标准资源库。资源通用数据模型（RCDM）。基于 RCDM，通过构建标准资源库，实现对不同来源、不同系统、不同结构、不同格式的原始业务数据的标准化治理，形成符合国家和行业标准的、统一的真实世界大数据标准资源，实现数据存储标准化。

1. 数据治理流程

（1）数据映射

数据映射是数据集成与融合的初始步骤，核心是建立源与目标数据模型间的元素级映射关系。基于业务镜像数据，系统依据通用标准模型进行抽取与转换，将原始数据直接对应到通用模型的指定表及字段。字段构建原则上采用直接映射，避免复杂统计计算，以保持数据的原始性与真实性。

（2）数据清洗

数据清洗是治理的首要环节，针对单行单字段数据进行处理，解决不同业务系统间数据格式不统一的问题。为满足通用数据模型对格式、长度、值域及内容的定义，系统进行以下三类清洗操作：

格式转换：将各医疗机构差异化的日期、数值、字符、布

尔及枚举等数据类型统一转换为标准格式。

通用计算：将基础逻辑计算前置，减轻后续分析负担，例如根据出生日期计算年龄、根据药品规格和数量计算标准化用药剂量。

字典映射：将机构自定义的枚举数据（如性别、婚姻状态代码）映射为平台统一的标准代码。

2. 数据质控

标准资源库数据质控是基于业务逻辑的专项质量控制，核心是验证数据采集的业务合理性与逻辑一致性，而非仅做技术层面的完整性检查。它通过关联不同业务表的内在关联关系，校验采集表、字段及数据内容的准确性，及时发现漏采、错采、逻辑矛盾等问题，保障资源库数据真实反映临床业务实际。

开展标准资源库业务逻辑质控，需先梳理各业务模块的核心关联规则，建立多表交叉校验指标体系，例如比对门诊就诊人次与诊断记录数的匹配比例、住院患者与医嘱记录的对应关系等；定期运行校验脚本生成异常数据清单，对低比例、逻辑矛盾的数据人工复核溯源，定位采集流程问题并整改，同时持续更新校验规则以适配业务变化。

（八）构建观察数据库

观察数据库是将疾病、药品、手术、检验等数据进行标准化处理，并清洗包括入院记录、出院小结、影像报告等非结构化文本数据，实现数据的归一化、标准化。

1. 数据治理流程

（1）数据标准化

疾病、药品、手术、检验四大编码数据标准化。将疾病、药品、手术、检验编码按照《国家临床版 2.0 诊断编码(ICD-10)》《国家临床版 3.0 手术操作编码（ ICD-9-CM3 ）》《国家药品统一编码 YPID》《医学实验室互认字典（ CNAS-AL09 ）》等标准完成编码归一化、标准化治理。如遇国家编码更新，项目需按照新标准更新。

（2）文本结构化

非结构化文本数据清洗。对于非结构化文本数据，例如入院记录、出院小结等，由医学知识组通过分析其段落标志，定义文档预处理规则，准确地拆分出文档中的段落信息。

2. 数据质控

观察数据库质控是针对疾病、药品、手术、检验数据标准化及非结构化文本结构化全流程的专项质量控制，核心是验证编码映射的准确性与文本解析的正确性，排查编码错配、漏标及文本拆分错误、信息提取不全等问题，保障数据归一化、标准化的质量，为临床观察研究提供可信基础。

开展观察数据库质控，需分模块建立校验规则：编码标准化方面，逐类比对原始数据与国家临床版编码、统一药品编码等国家标准的映射关系，定期抽查编码准确率，同步跟踪国家编码更新后的落地情况；文本结构化方面，验证文档段落拆分规则的合理性，抽查入院记录、出院小结等关键信息提取的完整度与正确性，发现问题及时优化治理规则并返工修正。

（九）构建科学变量库

科学变量库涵盖社会/经济/环境/地理、疾病、手术、药

品、检验、术语及中医症候等业务域，超过 10 万+数据变量，旨在统一定义和维护科学数据队列等所用变量，并提供详细变量指标说明。

1. 数据治理流程：

（1）创建变量集、变量组和变量

通过选择标准术语集及变量组模板，并用标准术语替换模板中的占位符，构建成变量组实例，同时生成变量实例、数据元实例和变量加工规则实例，并绑定科学变量与数据元的引用关系。在生成变量组实例后，选择科学数据队列仓库所需的变量组数据，生成其科学变量集。

（2）数据变量构建及治理

依托科学变量库开展数据变量全流程构建与治理，涵盖定义、分类、审核、更新，保障变量准确、一致、可用。研究人员按需选用组合变量，严格审核验证其定义、取值与算法，定期维护更新并合规管控，为临床研究提供可靠数据支撑，提升医疗数据价值与应用水平。

2. 数据质控

科学变量库质控是针对多业务域变量创建及全生命周期治理的专项质量控制，核心是验证超 10 万级变量的定义规范性、关联关系准确性与生成逻辑合理性，排查变量错配、引用混乱、加工规则错误等问题，保障变量库的统一性、一致性和可用性，为科学数据队列研究提供标准化变量基础。

开展科学变量库质控，需分环节建立校验体系：变量创建环节，检查标准术语替换、模板应用的规范性，验证变量与数

据元的绑定关系及实例生成逻辑；变量治理环节，审核变量定义、分类、取值与加工算法，定期抽查变量可用性，跟踪更新维护情况，建立多级审核机制，确保变量符合研究需求与标准规范。

（十）构建科学数据库

科学数据库（SCDM）按科学通用数据模型组织存储，基于科学变量库标准，面向宏观主题、涵盖全局患者/人群样本，是纵向数据队列仓库。

它以全域人群为对象建纵向队列，变量源自科学变量库，数据来自观察数据库，且按“患者标识、记录时间 t 、发生时间 t 、变量 $X_1, X_2, \dots, X_m$ ”纵向排列存储。

SCDM 需覆盖社会/经济/环境/地理、疾病等多核心数据域，构建全面标准化元数据变量体系，为流行病学研究、临床试验及真实世界研究提供数据支撑。

1. 数据治理流程：

（1）变量加工

将变量加工规则生成为完整的查询、取值和值域转换的脚本，以便在观察数据中能够一次性找到所需要的数据；而不需要在逻辑代码中做数据筛选和转换。

（2）数据转换

通过定义科学变量集，创建数据转换任务，根据上述变量加工规则，从观察数据库中获取科学变量集中所定义的所有变量组的变量数据，映射到科学数据任务库中。

（3）生成科学数据队列仓库

通过上述数据转换任务，生成科学数据后，可查看科学数据队列仓库的数据概况，包括各个变量与标准资源库、观察数据库、科学变量库的血缘关系等。可通过科学数据队列仓库可视化工具，浏览科学数据详情，以纵向数据队列的形式，展示所有科学变量组及变量值，也可查看其元数据格式。

2. 数据质控

科学数据库（SCDM）质控是针对纵向数据队列仓库全构建流程的专项质量控制，核心是验证变量加工逻辑、数据转换映射及队列生成的准确性，排查脚本错误、数据丢失、结构异常及血缘断裂等问题，保障多域数据按通用模型规范存储，为各类临床研究提供可信的纵向队列数据基础。

开展科学数据库质控，需按构建流程分阶段实施：变量加工环节，校验查询、取值及值域转换脚本是否符合科学变量库标准；数据转换环节，比对转换前后的数据量与变量完整性，验证患者标识、双时间维度等核心字段的准确性；队列生成环节，检查纵向存储结构规范性，追溯全链路数据血缘，抽查变量值与上游库的一致性，同时验证元数据体系及可视化展示的正确性。

三、安全与隐私保护

本章旨在确立真实世界研究（RWS）必须遵循的伦理底线与安全红线。鉴于“三医”数据涉及高度敏感的个人健康信息，项目组必须建立贯穿数据采集、治理、存储及分析全流程的保障机制，在挖掘数据价值的同时，确保护理对象的隐私权利不受侵犯，数据资产的安全处于可控状态。

（十一）隐私保护技术

1. 不可逆脱敏与受控追溯：为了保护个人隐私，保护数据代表的对象利益不受侵害，系统必须对所有入库数据进行脱敏处理。指南明确规定，所有的脱敏操作原则上均应是不可逆的，即数据一旦脱敏，在常规分析环境中严禁进行反脱敏操作。必须依据《中华人民共和国个人信息保护法》及 HIPAA 标准，移除或泛化姓名、身份证号等 18 类直接标识符。然而，考虑到临床安全事件核查的需求，对于患者信息，在个人隐私信息脱敏后，仅允许通过特定的物理隔离机制进行反向追溯。该追溯权限需实行最高级别的审批管理，严禁向普通分析人员开放。

2. 隐私计算技术的应用：针对部分极度敏感或无法出域的数据，如涉及人类遗传资源或传染病数据，应当采用隐私计算技术，特别是联邦学习。坚持“数据不动模型动，数据可用不可见”的原则，仅在各机构本地进行模型训练并交换加密梯度参数，而不交换原始数据。这种机制确保了在不泄露原始数据隐私的前提下，实现跨机构数据的联合分析与建模。

（十二）数据安全运维

1. 分级分类存储与隔离：数据中心应当实行严格的数据分级分类存储策略。对于经过脱敏处理的普通科研数据，可存放在分析区供授权人员使用；对于包含身份信息的密钥数据或极高敏感度的基因组数据，必须存储在物理隔离的绝密区，并采用高强度的加密算法进行存储加密。同时，应建立网络安全隔离区，严格控制内网与外网的数据交换，防止外部攻击与内部

泄露。

2. 细粒度权限控制与全链路审计：系统必须实施基于角色的访问控制，遵循最小够用原则分配数据访问权限。分析师仅能访问其研究课题所必需的最小数据集，严禁越权访问其他项目数据。此外，必须建立全链路审计日志机制，系统应自动记录所有用户的数据访问、查询、导出、修改等操作行为。审计日志应至少留存 5 年，确保任何数据操作均可追溯、可核查，一旦发生安全事件，能够迅速定位责任人并还原事件过程。

四、术语与定义

（十三）真实世界研究（RWS）

指在真实临床、医保、公共卫生等常规医疗服务环境中，利用真实世界数据开展的关于医疗干预措施的使用情况、疗效、安全性及经济性等方面的研究，其研究设计通常包括队列研究、病例对照研究、横断面研究、注册研究等，区别于传统的随机对照试验（RCT）。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（真实世界数据）或基于这些数据衍生的汇总数据，通过分析，获得药物的使用情况及潜在获益-风险的临床证据（真实世界证据）的研究过程。包括但不限于干预性研究、非干预性研究或观察性研究（包括前瞻性研究，回顾性研究）。

（十四）真实世界数据（RWD）

指在日常医疗照护、公共卫生管理、医保支付等非试验性场景中常规收集的，与个体健康状况、医疗服务利用、诊疗过程及结局等相关的数据，包括但不限于电子病历、医保结算数

据、疾病登记数据、药品不良反应监测数据、可穿戴设备数据等。指从常规医疗服务过程中收集的数据，例如，健康居民档案、电子病历信息、医院收费数据、医保支付数据、疾病登记数据、不良反应监测数据、死亡登记数据等。

（十五）真实世界证据（RWE）

指通过对真实世界数据进行系统采集、清洗、标准化和统计分析后，所获得的关于医疗产品使用模式、疗效、安全性和经济性的临床或决策证据，可用于支持药物和医疗器械的研发、监管决策、医保支付政策制定以及临床实践指南更新等。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，包括通过对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。

（十六）通用数据模型（CDM）

指为实现不同来源、不同结构的健康数据在语义和结构上的统一，而制定的一套标准化数据模型和编码规范。它通常包括统一的实体、数据结构和术语标准，用于支持多中心数据整合、跨机构分析和真实世界研究。常见的 CDM 包括 OMOP CDM、PCORnet CDM 等。

（十七）JDBC (Java Database Connectivity)

是一种用于 Java 编程语言的应用程序接口（API），定义了 Java 程序与各种关系型数据库进行连接、查询和数据操作的标准方法。通过 JDBC，应用程序可以统一访问不同数据库管理系统（如 MySQL、Oracle、SQL Server 等），实现数据库无

关性和数据访问的标准化。

(十八) 患者主索引 (Enterprise Master Patient Index, EMPI)

指在一个机构或跨多个机构范围内，用于唯一识别和管理患者身份信息的核心索引系统。它通过算法和规则（如姓名、身份证号、出生日期、性别等信息的匹配）对来自不同系统的患者记录进行去重、合并和关联，形成统一、唯一的患者标识，以支持医疗数据整合、患者连续管理和跨系统数据共享。