用于产生真实世界证据的真实世界数据
指导原则
（征求意见稿）

2020 年 7 月

目 录

用于产生真实世界证据的真实世界数据指导原则

一、概述

真实世界证据（Real World Evidence，RWE）是药物有效性和安全性评价证据链的重要组成部分，其相关概念和应用参见《真实世界证据支持药物研发与审评的指导原则（试行）》。而真实世界数据（Real World Data, RWD）则是产生RWE的基础，没有高质量的适用的RWD支持，RWE亦无从谈起。

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/或诊疗及保健有关的数据。并非所有的真实世界数据经分析后就能产生真实世界证据，只有满足适用性的RWD经恰当和充分地分析后才有可能形成RWE。目前RWD普遍存在数据的记录、采集、存储等流程缺乏严格的质量控制，数据不完整，数据标准和数据模型不统一等问题，对RWD的有效使用形成了障碍。因此，如何使收集的RWD能够成为或经治理后能够成为满足临床研究目的所需的分析数据，以及如何评估RWD是否适用于产生RWE，是使用真实世界数据形成真实世界证据支持药物监管决策的关键问题。

本指导原则作为《真实世界证据支持药物研发与审评的指导原则（试行）》的补充，将从真实世界数据的定义、来

源、评价、治理、标准、安全合规、质量保障、适用性等方面，对真实世界数据给出具体要求和指导性建议，以帮助申办者更好地进行数据治理，评估 RWD 的适用性，为产生有效的 RWE 做好充分准备。

二、 真实世界数据来源及现状

与药物研发有关的真实世界数据主要包括在真实医疗环境下业务流程记录的数据（如电子病历），以及各种观察性研究数据等。此类数据可以是开展真实世界研究前已经收集的数据，也可以是为了开展真实世界研究而新收集的数据。

（一）真实世界数据常见的主要来源

我国真实世界数据的来源按功能类型主要可分为医院信息系统数据、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，以下是根据数据功能类型分类的常见真实世界数据来源。

1. 医院信息系统数据

医院信息系统数据包括结构化和非结构化的数字化或非数字化患者记录，如患者的人口学特征、临床特征、诊断、治疗、实验室检查、安全性和临床结局等，通常分散存储于医疗卫生机构的电子病历/电子健康档案（EMR/EHR）、实验室信息管理系统（LIS）、医学影像存档与通讯系统（PACS）、

放射信息管理系统（RIS）等不同信息系统中。有些医疗机构在数据集成平台或临床数据中心（CDR）的基础上建立院级科研数据平台，整合患者门诊、住院、随访等各类信息，形成直接用于临床研究的数据。有些区域性医疗数据库，利用相对集中的物理环境进行跨医疗机构的临床数据的存储和处理，具有存储量大、类型多等特点，也可作为 RWD 的潜在来源。

医院信息系统数据基于临床诊疗实践过程的记录，涵盖临床结局和暴露变量范围较广，尤其电子病历数据在真实世界研究中应用较广。

2. 医保支付数据

我国医保支付数据的主要来源有两类，一类是政府、医疗机构建立的基本医疗保险体系，进行医保支付数据库的建立和统一管理，包含有关患者基本信息、医疗服务利用、处方、结算、医疗索赔和计划保健等结构化字段的数据。另一类是商业健康保险数据库，由保险机构建立，数据以保险公司理赔给付与保险期限作为分类指标，数据维度相对简单。医保系统作为真实世界数据来源，较多用于开展卫生技术评价和药物经济学研究。

3. 登记研究数据

登记研究（Registry Study）数据是通过有组织的系统，

利用观察性研究的方法搜集临床和其他来源的数据，可用于评价特定疾病、特定健康状况和暴露人群的临床结局。登记研究根据研究定义的人群特点主要包括产品登记、健康服务登记和疾病登记三类，我国的登记研究主要是疾病登记和产品登记研究。其中，医疗机构和企业支持开展的产品登记研究，观察对象是使用某种医药产品的病例，重点观察其不同适应症的效果或监测不良反应。

登记研究数据库的优势在于以特定患者为研究人群，通过整合临床诊疗、医保支付等多种数据来源，数据采集较为规范，一般包括患者自报数据和长期随访数据，观测结局指标通常较为丰富，具有准确性较高、结构化强、人群代表性较好等优点，对于评价药物的有效性、安全性、经济性和依从性具有较好的适用性。

4. 药品安全性主动监测数据

药品安全性主动监测数据主要用于开展药物安全性研究及药物流行病学研究，通过国家或区域药品安全性监测网络，从医疗机构、制药公司、医学文献、网络媒体、患者报告结局等渠道，进行数据收集。此外，医疗机构和企业自身建立的自有药品的安全性监测数据库也可能成为此类数据来源的一部分。

5. 自然人群队列数据

自然人群队列数据指对普通人群或患有重大疾病人群通过长期前瞻性动态追踪观察，获取的各种数据。自然人群队列数据具有统一标准、信息化共享、时间跨度长和样本量较大的特点，此类 RWD 可以帮助构建常见疾病风险模型，可对药物研发的精准目标人群定位提供支持。

6. 组学数据

组学数据作为精准医学的重要支撑，主要包括基因组、表观遗传、转录组、蛋白质组和代谢组等数据，这些数据从系统生物学角度刻画了患者在遗传、生理学、生物学等方面的特征。通常组学数据需要结合临床数据才可能成为适用的 RWD。

7. 死亡登记数据

人口死亡登记是一个国家对其国民的死亡信息持续完整的收集和记录。目前我国有四个系统用于收集人口死亡信息，分别隶属于国家疾控中心、国家卫生健康委员会、公安部和民政部。人口死亡登记数据包含死亡医学证明书中的所有信息，记录了详细的死亡原因和死亡时间，可以产出人群分死因死亡率的数据来源。

8. 患者报告结局数据

患者报告结局（Patient-Reported Outcome, PRO）是一种来自患者自身测量与评价疾病结局的指标，包括症状、生理、

心理、医疗服务满意度等，PRO 在药物评价体系发展中越来越重要。其记录有纸质和电子两种方式，后者称为电子患者报告结局（ePRO），ePRO 的兴起与应用，使得 PRO 与电子病历系统对接并形成患者层面的完整数据流成为可能。

9. 来自移动设备的个体健康监测数据

个人健康监测数据可通过移动设备（如智能手机、可穿戴设备）实时采集个体生理体征指标。这些数据常产生于普通人群的自我健康管理、医疗机构对慢病患者的监测、医疗保险公司对参保人群健康状况评估的过程，通常存储于可穿戴设备企业、医疗机构数据库以及商业保险公司数据系统等。由于可穿戴设备在收集生理和体征数据方面具有便利性和即时性等优势，与电子健康数据衔接可形成更完整的 RWD。

10. 其它特定功能数据

（1） 公共卫生监测数据

我国建立了一系列有关公共卫生监测的数据库，如传染病监测、免疫接种不良事件(Adverse Events Following Immunization，AEFI)监测等，所记录的数据可用于分析传染病的发病情况、疫苗的一般反应和异常反应发生率等。

（2） 患者随访数据

在真实世界临床诊疗环境中，院内电子病历数据往往无法涵盖患者一些重要的临床指标，如总生存期、五年生存率、

不良反应信息等，需要补充长期随访数据，才能形成适用的RWD。患者随访数据主要是指以临床研究为目的，医院随访部门或第三方授权服务商以信件、电话、门诊、短信、网络随访等方式对离院患者开展临床终点、康复指导、用药提醒、满意度调查等服务，服务中收集的院外数据，通常存储于医院随访数据系统。通过与病历数据的连接，实现多源临床数据的融合，可最终形成覆盖患者生命周期的完整数据，用以探索疾病发生机制、发展规律、治疗方法、预后相关因素等临床研究问题。

（3） 患者用药数据

患者诊疗过程药品使用数据包括患者信息、药品品类、剂量以及不良反应等信息，通常存储于医院药品管理信息系统、医药电子商务平台、制药企业产品追溯和药品安全性信息数据库，以及药品使用监测平台等。伴随远程诊疗和互联网+慢病管理模式的普及，存储于处方流转平台或医药电商平台的患者院外用药数据逐渐增多，此类数据的有效利用或拼接，可作为患者维度诊疗过程记录的RWD来源。

随着医疗信息技术的不断发展，新的RWD类型和来源会不断出现，但其具体应用还有赖于所要解决的临床研究问题，以及该数据所支持产生RWE的适用性。

（二）真实世界数据应用面临的主要问题

从数据来源看，相较于RCT数据，RWD在大多数情况下缺乏其记录、采集、存储等流程的严格质量控制，会造成数据不完整、关键变量缺失、记录不准确等问题，这些数据质量上的缺陷，会极大地影响后续的数据治理和应用，甚至会影响数据的可追溯性，研究者也难以发现其中的问题并进行核对和修正。由于患者病程、就诊地点以及时间和空间等因素的变化，可能导致患者疾病状态及相关因素等信息的缺失，为临床研究疾病状态及结局的系统性评价带来挑战。倾向性的数据收集，特别是登记研究数据，会导致研究结果偏倚的潜在风险。

由于各种RWD来源之间相对独立和封闭、数据管理系统种类繁多、数据存储分散且数据标准不一致、数据横向整合和交换存在困难，造成数据碎片化和信息孤岛现象突出。对于电子病历数据，由于其高度敏感性，业务系统一般封闭管理，对它们的利用可能会受到一定限制。此外，在缺乏统一标准的情况下，数据类型较为多样，既有结构化数据，也有文本、图片、视频等非结构化和半结构化数据，在数据记录、采集、存储的过程中，也会导致数据的冗余和重复，进而造成数据处理难度加大。

三、 真实世界数据适用性评价

真实世界数据的适用性评价应基于特定的研究目的和

监管决策用途。适用性评价可分为两个阶段，第一阶段是从可及性、伦理合规、代表性、关键变量完整性、样本量和源数据活动状态等维度，对源数据进行初步评价和选择，判断其是否满足研究方案的基本分析要求；第二阶段包括数据的相关性、可靠性，以及采用的或拟采用的数据治理机制（数据标准和通用数据模型）的评价分析，经治理的数据是否适用于产生真实世界证据（见图 1）。如果真实世界研究中研究者根据自己设计好的电子病例报告表（eCRF）前瞻性收录指定来源数据，则无需进行第一阶段的初步适用性评价。

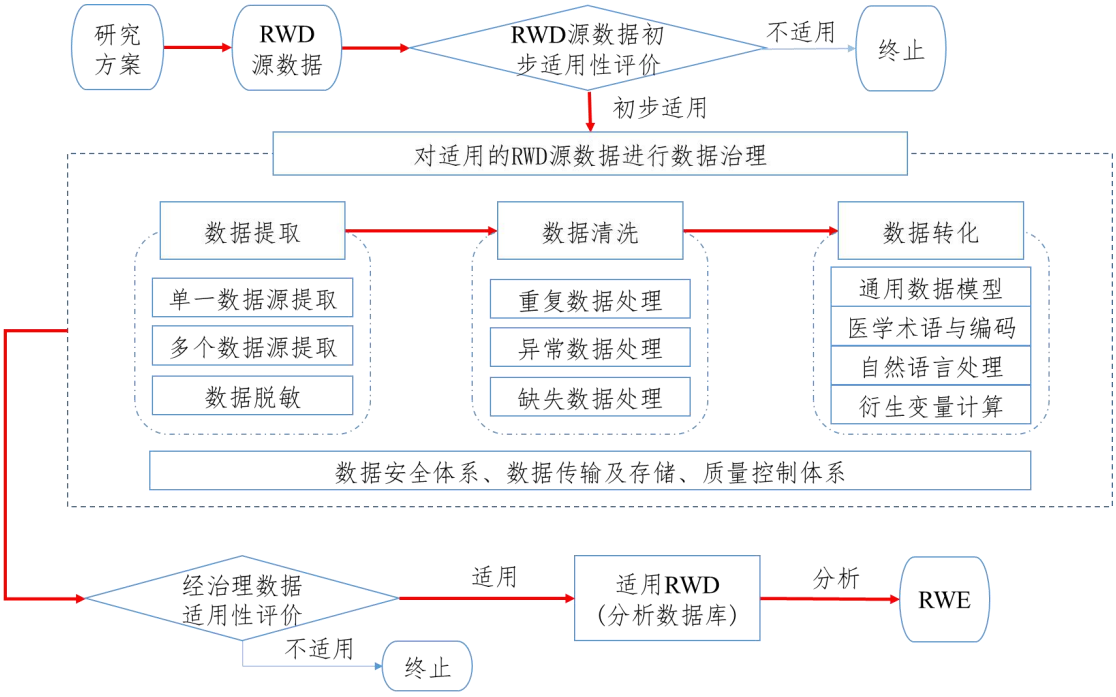

图 1 真实世界数据的适用性评价和数据治理过程示意图

（一）源数据的适用性评价

满足基本分析要求的源数据至少应具备以下条件：

1. 数据库处于活动状态且数据可及

在研究期间数据库应是连续的处于活动状态的，所记录的数据均是可及的，即具有数据的使用权限，并且可被第三方特别是监管机构评估。

2. 符合伦理和数据安全性要求

源数据的使用应通过伦理审核，并符合数据安全性要求。

3. 临床结局和暴露/干预变量

数据的记录必须有临床结局变量和暴露/干预变量。

4. 具有一定的数据完整性

源数据通常是不完整的，但应具有一定的完整性，至少应包括结局变量、暴露/干预变量、人口学变量和重要的协变量，还要考虑分析模型中缺失数据对研究结论稳健性的影响。

5. 样本量足够

应充分考虑和预判经数据治理后源数据例数明显减少的情况，以保证统计分析所需的样本量。

(二) 经治理数据的适用性评价

经治理的真实世界数据的适用性评价主要根据数据相关性和可靠性。

1. 相关性评价

相关性评价旨在评估 RWD 是否与所关注的临床问题密切相关，重点关注关键变量的覆盖度、临床结局定义的准确性、目标人群的代表性和多源异构数据的融合性。

（1） 关键变量和信息的覆盖度

RWD 应包含与临床结局相关的重要变量和信息，如药物使用、患者人口学和临床特征、协变量、结局变量、随访时间、潜在安全性信息等。如果上述变量存在部分缺失，需充分评估是否能够使用可靠的估计方法进行填补，以及对于因果推断结果可能造成的影响。

（2） 临床结局定义的准确性

选择具有临床意义的结局并进行准确定义对于真实世界研究至关重要。临床结局的定义应包括所基于的诊断标准、测量方法及其质量控制（如果有）、测量工具（如量表的使用）、计算方法、测量时点、变量类型、变量类型的转换（如从定量转换为定性）、终点事件评价机制（如终点事件委员会的运行机制）等。当不同数据源对临床结局的定义不一致时，应定义统一的临床结局，并采用可靠的转换方法。

（3） 目标人群的代表性

真实世界研究较传统 RCT 的优势之一是具有更广泛的目标人群的代表性。因此，在制定纳入和排除标准时，应尽可能地符合真实世界环境下目标人群。

（4） 多源异构数据的融合性

由于 RWD 的特性，很多情况下属于多来源的异构数据，需要将不同来源数据在个体水平进行数据的链接、融合和同

构处理。因此，应通过身份标识符进行个体水平的准确链接，以支持通用数据模型或数据标准对数据源中关键变量进行整合。

2. 可靠性评价

真实世界数据的可靠性主要从数据的完整性（Completeness）、准确性（Accuracy）、透明性（Transparency）、质量控制（Quality Control）和质量保证（Quality Assurance）几个方面进行评价。

(1) 完整性

完整性是指数据信息的缺失程度，包括变量的缺失和变量值的缺失。对于不同研究，数据的缺失程度、缺失原因和变量值的缺失机制不尽相同，应该予以详尽描述。虽然 RWD 无法避免数据缺失问题，但缺失比例应有一定限度。当特定研究的数据缺失比例明显超过同类研究的比例时，会加大研究结论的不确定性，此时需要慎重考虑该数据能否作为支持产生 RWE 的数据。对缺失原因的详细分析有助于对数据可靠性的综合判断。如果涉及缺失数据的填补问题，应根据缺失机制的合理假设采用正确的填补方法。

(2） 准确性

准确性是指数据与其描述的客观特征是否一致，包括源数据是否准确、数据值域是否在合理范围、结局变量随时间

变化趋势是否合理、编码映射关系是否对应且唯一等。数据的准确性需要依据较权威的参照进行识别和验证，例如，终点事件是否经独立的终点事件委员会做出判断。

（3） 透明性

RWD的透明性是指RWD的治理方案和治理过程清晰透明，应确保关键暴露变量、协变量和结局变量能够追溯至源数据，并反映数据的提取、清洗、转换和标准化过程。无论采用人工数据处理还是自动化程序处理，数据治理标准化操作程序和验证确认文件要清晰记录和存档，尤其反映数据可信性的问题，如数据缺失度、变量阈值范围、衍生变量计算方法和映射关系等。数据治理方案应事先根据研究目的制定，应确保数据治理过程与治理方案保持一致。数据的透明性还包括数据的可及性（Accessibility）、数据库之间的信息共享和对患者隐私的保护方法的透明。

（4） 质量控制

质量控制是指用以确证数据治理的各个环节符合质量要求而实施的技术和活动。质量控制评价包括但不限于：数据提取、安全处理、清洗、结构化，以及后续的存储、传输、分析和递交等环节是否均有质量控制，以保证所有数据是可靠的，数据处理过程是正确的；是否遵循完整、规范、可靠的数据治理方案和计划，并依托于相应的数据质量核查和系

统验证规程，以保障数据治理系统在正常和稳态下运行，确保真实世界数据的准确性和可靠性。

（5） 质量保证

质量保证是指预防、探测和纠正研究过程中出现的数据错误或问题的系统性措施。RWD 的质量保证与监管合规性密切相关，应贯穿于数据治理的每一个环节，考虑的内容包括但不限于：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；是否有相应的标准操作规程；数据收集是否有明确流程和合格人员；是否使用了共同的定义框架，即数据字典；是否遵守收集关键数据变量的共同时间框架；用于数据元素捕获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、与保险数据的链接等；患者的选择是否将偏倚最小化以体现真正的目标人群；数据输入是否及时、传输是否安全；是否满足监管机构现场核查调阅源数据、源文件等相关要求。

四、真实世界数据治理

数据治理（Data Curation）是指针对特定临床研究问题，为达到适用于统计分析而对原始数据所进行的治理，其内容包括但不限于：数据安全性处理、数据提取（含多个数据源）、数据清洗（逻辑核查及异常数据处理、数据完整性处理）、数据转化（数据标准、通用数据模型、归一化、自然语言处

理、医学编码、衍生变量计算）、数据传输和存储、数据质量控制等若干环节。

（一）个人信息保护和数据安全性处理

真实世界研究涉及个人信息保护应遵循国家信息安全技术规范、医疗大数据安全管理相关规定，对个人敏感信息应进行去标识化（de-identification）处理，确保根据数据无法进行个人敏感信息匹配还原，通过技术和管理方面的措施，防止个人信息的泄漏、损毁、丢失、篡改。

数据安全性处理应基于研究所涉及的各种数据的类型、数量、性质和内容，尤其对于个人敏感信息，建立数据治理各环节的数据加密技术要求、风险评估和应急处置操作规程，并开展安全措施有效性审计。

（二）数据提取

根据源数据的存储格式、是否为电子数据、是否包含非结构化数据等因素选择合适的方式进行数据提取，在数据提取时均应遵守以下原则：

数据提取的方法应通过验证，以保障提取到的数据符合研究方案的要求。数据提取应确保提取到的原始数据与源数据的准确性，应对提取到的原始数据与源数据进行时间戳管理。

使用与源数据系统可互操作或集成的数据提取工具可

以减少数据转录中的错误，从而提高数据准确性以及临床研究中数据采集的质量和效率。对于盲法研究，还应评估使用可互操作或集成的数据提取工具带来的揭盲风险。

（三）数据清洗

数据清洗（Data Cleaning）是指对提取的原始数据进行重复或冗余数据的去除、变量值逻辑核查（Edit Check）和异常值的处理，以及数据缺失的处理。需要注意，在修正数据时如果无法追溯到主要研究者或源数据负责方签字确认，数据不应做修改，以保证数据的真实性。

首先在保证数据完整性的前提下去除重复数据及不相关数据。在不同数据源合并过程中，可能产生重复数据，需要去除。同时由于数据源与通用数据模型映射关系的不准确，可能会采集到与研究目标不相关的数据，从数据集中删除不需要的观测值可以减少不必要的工作。

然后进行逻辑核查和异常数据处理。通过逻辑核查可以发现原始数据或者提取数据时产生的错误，例如出院时间早于入院时间，出生年月按年龄推算不符，实验室检查结果不符合实际，定性判断结果与方案中定义的判断标准不一致等。对异常数据的处理要非常谨慎，避免由此产生的偏倚。对于发现的错误和异常数据应通过进一步核实才能更改数据，数据的更改应保留记录。

最后对数据缺失进行处理，对于不同研究，数据的缺失程度、缺失原因和变量值的缺失机制不尽相同。如果涉及缺失数据的填补问题，应根据缺失机制的合理假设采用正确的填补方法。

（四）数据转化

数据转化是将经过数据清洗后原始数据的数据格式标准、医学术语、编码标准、衍生变量计算，按照分析数据库（Analysis Dataset）中对应标准进行统一转化为适用 RWD 的过程。

对于自由文本数据的转化可使用可靠的自然语言处理算法，在保障数据转化准确、可溯源的前提下，提高转化效率。

在进行衍生变量计算时，应明确用于计算的原始数据变量及变量值、计算方法及衍生变量的定义，并进行时间戳管理，以保障数据的准确性和可追溯性。

（五）数据传输和存储

真实世界数据的传输和存储应当基于可信的网络安全环境，在数据收集、处理、分析至销毁的全生命周期予以控制。在数据传输和存储过程中都应有加密保护。此外，应建立操作设置审批流程、角色权限控制和最小授权的访问控制策略，鼓励建立自动化审计系统，监测记录数据的处理和访

问活动。

（六）数据质量控制

数据质量控制是确保研究数据完整性、准确性和透明性的关键。数据质量控制需要建立完善的RWD质量管理体系和SOP，建议原则包括:

1. 确保源数据的准确性和真实性

如电子病历作为关键数据源，应有病历质控标准以满足分析要求。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。

2. 在数据提取时充分考虑数据完整性问题

评估和确立提取字段，制定相应的核查规则和数据库架构。

3. 建立数据录入和结构化的标准指南，确保录入数据与源数据的一致性。

对于录入过程中的任何修改，需要有负责人的确认和签名，并提供修改原因，确保留下完整的稽查轨迹。

4. 制定完善的数据质量管理计划

制定系统质控和人工质控计划，确保数据的准确性和完整性。对于关键变量，应进行100%核查和源文件调阅；其它变量可根据实际情况抽样核查，例如，对于人口学信息、数值型变量阈值、编码映射关系等，可按一定比例抽样，核

查其准确性与合理性。

（七）通用数据模型

通用数据模型（Common Data Model，CDM）是多学科合作模式下对多源异构数据进行快速集中和标准化处理的数据系统，其主要功能是将不同标准的源数据转换为统一的结构、格式和术语，以便跨数据库/数据集进行数据整合。

由于多源数据的结构和类型的复杂性、样本规模和标准的差异性，在将源数据转换为 CDM 的整体过程中，需要对源数据进行提取、转换、加载（Extract-Transform-Load，ETL），应确保源数据在语法和语义上与目标分析数据库的结构和术语一致。

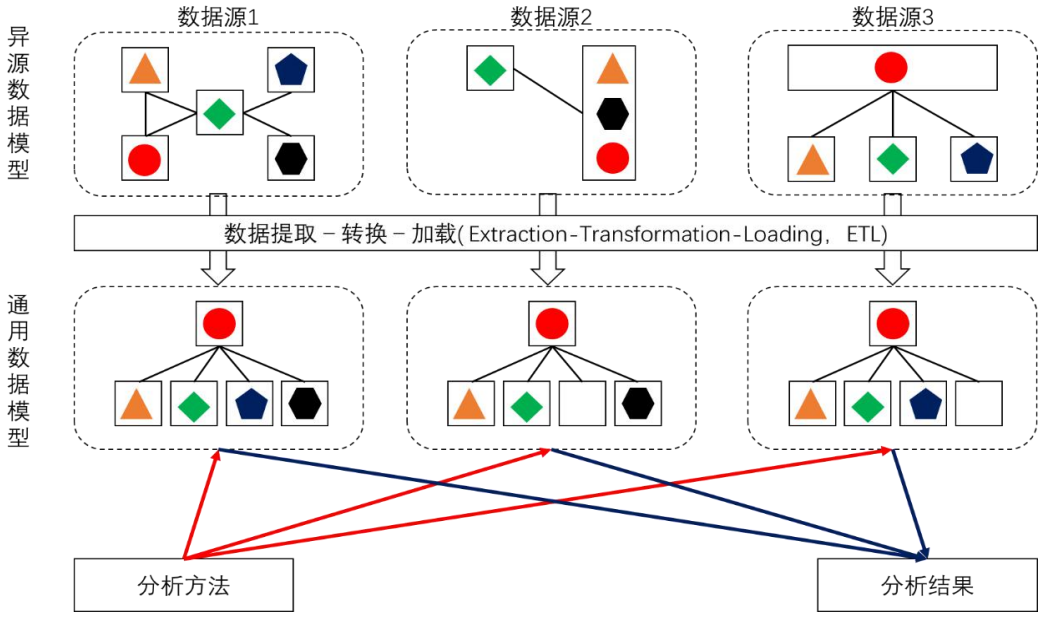

图 2 异源数据模型向通用数据模型转化的示意图

理想的通用数据模型应遵循以下原则：

1. CDM 可以定义为一种数据治理机制，通过该机制可以

将源数据标准化为通用结构、格式和术语，从而允许跨多个数据库/数据集进行数据整合。CDM 应具有访问源数据的能力，是可动态扩展和持续改进的数据模型，并有版本控制；

2. CDM 变量的定义、测量、合并、记录及其相应的验证应保持透明，多个数据库的数据转换应有清晰一致的规则；

3. CDM 应有基线概念，安全性和有效性相关的常用变量或概念都应映射到 CDM，以适用于不同临床研究问题，并可通过公认或已知的研究结果进行比对。

（八）真实世界数据治理计划书

RWD 治理计划书应事先制定，与整个项目研究计划同步。如果治理计划书在研究进行过程中需要修订，应与审评机构沟通并备案。计划书中应说明使用 RWD 用于监管决策的目的、使用 RWD 的研究设计，还应对 RWD 源数据进行说明，包括但不限于：RWD 源数据/源文件的类型，例如卫生信息系统数据、疾病登记数据、医保数据等；RWD 的源数据/源文件，适当评价其既往应用情况，说明采用的理由；RWD 的治理，即由 RWD 数据来源到分析数据库的治理过程，包括数据提取、数据安全性处理、数据清洗（逻辑核查及异常数据处理、数据完整性处理）、数据导入和结构化（通用数据模型、归一化、自然语言处理、医学编码、衍生变量）、数据传输等；采用的数据模型和数据标准；缺失数据的处理

方法；减少或控制使用 RWD 带来的潜在偏倚所采取的措施；质量控制和质量保证；RWD 的适用性评估，即申报项目对 RWD 的适用性评估准则。

五、 真实世界数据的合规性、安全性与质量管理体系

（一）数据合规性

RWD 来源于患者个人诊疗等多种途径的数据，数据的收集、处理与使用等会涉及伦理及患者隐私问题。为充分保护患者的安全和权益，获取和使用 RWD 以开展真实世界研究，须通过伦理委员会的审核批准。参与 RWD 治理的相关人员需严格遵守相关法律、法规的要求，申办者应严格执行，尽保护和管理义务。

（二）数据安全管理

应依照国家法律法规、行业监管要求等做好数据安全管理工作，对承载健康医疗数据的信息系统和网络设施以及云平台等进行必要的安全保护。数据安全保护范围应涵盖包括数据收集、数据提取、数据传输、数据存储、数据交换、数据销毁等在内的各个生命周期。采用加密技术保证数据在收集、提取、传输和存储过程中的完整性、保密性、可追溯性，使用介质传输的，应对介质实施管控。对不同介质的数据形式采用不同的保护措施，并建立相对应的访问控制机制，对访问记录进行审核、登记、归档和审计。

数据审计及相关操作规程为数据的收集、提取、传输、维护、存储、共享、使用等提供记录和依据，应包括人员审计、管理审计、技术审计，应制定和部署医疗信息系统活动审计政策和适当的标准操作流程。审计的内容应包括数据的任何状态的任何操作，包括登录、创建、修改和删除记录的行为，都应自动生成带有时间标记的审计记录，包括但不限于授权信息、操作时间、操作原因、操作内容、操作人及签名等信息，并可供审计。审计记录应被安全存储并建立访问控制策略。

（三）质量管理体系

应建立完整的质量管理体系，以规范 RWD 的处理流程，并在实际工作中持续优化、完善。基本质量要素应覆盖：确保 RWD 的质量，应建立覆盖 RWD 全生命周期管理的操作流程；计算机化系统功能应满足 RWD 的管理需求，符合相关法规对计算机化系统的相关要求；建立完善的人员管理制度，数据收集、治理、分析人员应获得相应的培训，符合职责能力要求，并对人员的权限进行标准化管理；建立从数据收集至数据递交各环节的风险管理流程；制定标准的信息与文档管理规范（纸质、电子介质），确保 RWD 处理流程记录完整、准确、透明，保护数据的安全性与合规性。

六、与监管机构的沟通

为保证 RWD 的质量符合监管要求，鼓励申请人与监管机构及时沟通交流。在真实世界研究正式开始前，基于整体研发策略和具体研究方案等，就 RWD 是否支持产生 RWE 进行交流，包括 RWD 的可及性、样本量是否足够大、数据治理计划是否合理可行、数据质量可否得到保障等。在研究进行中，如果根据研究实施中的变化情况对数据治理计划进行调整，申办者需衡量数据治理计划调整对试验目标的潜在影响，向监管机构说明调整的充分理由，并征得其同意，还应将更新的研究方案和数据治理计划书备案。在研究完成后和递交资料前，申办者可与监管机构咨询递交资料和数据库进行沟通。

名词解释

电子病历（Electronic Medical Record，EMR)：由医疗机构中授权的临床专业人员创建、收集、管理和访问的个体患者的健康相关信息电子记录。

电子健康档案（Electronic Health Record，EHR）：符合国家认可的使用的互操作性标准，并能够由多个医疗机构中授权的临床专业人员创建、管理和咨询的针对个体患者的健康相关信息电子记录。

分析数据库（Analysis Dataset）：根据特定研究的具体要求，对原始数据库做该研究特有的定制化处理后形成的数据库，包括从研究中心提取原始数据补足缺失项、完成随访、通过患者 ID 进行数据关联、衍生指标计算、数据标准化、医学编码等。

观察性研究（Observational Study）：根据特定研究问题，不施加主动干预的、以自然人群或临床人群为对象的、探索暴露/治疗与结局因果关系的研究。

患者报告结局（Patient-Reported Outcome，PRO）：是一种来自患者自身测量与评价疾病结局的指标，包括症状、生理、心理、医疗服务满意度等。其记录有纸质和电子两种方式，后者称为电子患者报告结局（ePRO）。

逻辑核查（Edit Check）：对输入计算机系统的临床研究数据的有效性的检查，主要评价输入数据与其预期的数值逻辑、数值范围或数值属性等方面是否存在逻辑性错误。

数据标准（Data Standard）：是关于如何在计算机系统之间构建、定义、格式化或交换特定类型数据的一系列规则。数据标准可使递交的资料具有可预测性和一致性，且具有信息技术系统或科学工具可以使用的形式。

数据清洗（Data Cleaning）：数据清洗旨在识别和纠正数据中的噪声，将噪声对数据分析结果的影响降至最低。数据中的噪声主要包括不完整的数据、冗余的数据、冲突的数据和错误的数据等。

数据融合（Data Linkage）：将多来源的数据和信息加以合并、关联及组合，形成统一的数据集。

数据元素（Data Element）：临床研究中记录的受试者的单一观察值，例如，出生日期，白细胞计数，疼痛严重程度，以及其它临床观察值。

数据治理（Data Curation）:针对特定临床研究问题，为达到适用于统计分析而对原始数据所进行的治理，其内容至少包括数据提取（含多个数据源）、数据安全性处理、数据清洗（逻辑核查及异常数据处理、数据完整性处理）、数据转

化（通用数据模型、归一化、自然语言处理、医学编码、衍生变量计算）、数据质量控制、数据传输和存储等若干环节。

通用数据模型（Common Data Model，CDM）：是多学科合作模式下对多源异构数据进行快速集中和标准化处理的数据系统，其主要功能是将不同数据标准的源数据转换为统一的结构、格式和术语，以便跨数据库/数据集进行数据整合。

协变量（Covariate）：研究者预计的或通过探索性分析确定的会对主要结局变量产生重要影响的变量，它可以分为基线协变量和非基线协变量两类。

源数据（Source Data）：临床研究中记录的临床症状、观测值和用于重建和评估该研究的其他活动的原始记录和核证副本上的所有信息。源数据包含在源文件中（包括原始记录或其有效副本）。

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。

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。

真实世界证据（Real-World Evidence，RWE）：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。

参考文献

[1] 蔡婷, 詹思延. 加快我国疫苗安全主动监测系统建设的思考[J]. 中华预防医学杂志. 2019,53(7): 664-667.

[2] 国家卫生健康委, 国家药品监督管理局. 《药物临床试验质量管理规范》. 2020.07.01.

[3] 国家药品监督管理局药品审评中心.《临床试验数据管理工作技术指南》. 2016.07.27.

[4] 国家药品监督管理局.《真实世界证据支持药物研发与审评的指导原则（试行）》. 2020.01.07.

[5] 侯永芳, 宋海波, 刘红亮, 等. 基于中国医院药物警戒系统开展主动监测的实践与探讨[J]. 中国药物警戒. 2019,16(4): 212-214.

[6] 周莉, 欧阳文伟, 李庚, 等. 中国登记研究的现状分析[J] 中国循证医学杂志. 2019,19(6): 702-707.

[7] Berger M, Daniel G, Frank K, et al. A framework for regulatory use of real world evidence. https://healthpolicy.duke.edu/sites/default/files/atoms/files/rwe_white_paper_2017.09.06.pdf.

[8] Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care[J]. Nat Rev Clin Oncol. 2019,16(5): 312-325.

[9] Duke-Margolis Center for Health Policy. Characterizing RWD Quality and Relevancy for Regulatory Purposes. https://healthpolicy.duke.edu/publications.

[10] Duke-Margolis Center for Health Policy. Determining Real-World Data's Fitness for Use and the Role of Reliability. https://healthpolicy.duke.edu/publications.

[11] EMA. Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/reflection-paper-expectations-electronic-source-data-data-transcribed-electronic-data-collection_en.pdf.

[12] EMA. A Common Data Model for Europe – Why? Which? How? https://www.ema.europa.eu/en/documents/report/common-data-model-europe-why-which-how-workshop-report_en.pdf.

[13] Khozin S, Abernethy AP, Nussbaum NC, et al. Characteristics of real-world metastatic non-small cell lung cancer patients treated with nivolumab and pembrolizumab during the year following approval [J]. Oncologist. 2018, 23: 328–336.

[14] OHDSI – Observational Health Data Sciences and Informatics, https://www. ohdsi.org.

[15] Ong TC, Kahn MG, Kwan BM, et al. Dynamic ETL: a hybrid approach for health data extraction transformation and loading J.BMC Medical Informatics and Decision Making 2017, 17(1) : 134.

附录 1 中英文对照表

中英文词汇对照

中文	英文
标准操作规程	Standard Operation Procedure, SOP
病例报告表	Case Report Form, CRF
病例登记	Patient Registry
电子病历	Electronic Medical Record, EMR
电子健康档案	Electronic Health Record, EHR
分析数据库	Analysis Dataset
观察性研究	Observational Study
患者报告结局	Patient Reported Outcome, PRO
结局变量	Outcome Variable
可追溯性	Traceability
临床数据中心	Clinical Data Repository
逻辑核查	Edit Check
免疫接种不良事件	Adverse Events Following Mmunization, AEFI
数据标准	Data Standard
数据清洗	Data Cleaning
数据元素	Data Element
数据治理	Data Curation

中文	英文
通用数据模型	Common Data Model, CDM
卫生信息系统	Hospital Information System, HIS
适用真实世界数据	Research-ready RWD
衍生变量	Derived Variable
医保支付数据	Medical Claims Data
源数据	Source Data
真实世界数据	Real World Data, RWD
真实世界研究	Real World Research/Study, RWR/RWS
真实世界证据	Real World Evidence,RWE

576