·方法学指南解读·

《药物流行病学研究方法学指南（第2版）》系列解读（19）：通用数据模型数据分析的工具体系与方法学范式

吴昀效[1,2]，聂晓璐[3]，郑泳祺[1,2]，王聪慧[4]，吴嘉瑞[5]，孙烨祥[6]，宋海波[7]，詹思延[1,2,8,9]，孙　凤[1,2,8,10,11,12]

1. 重大疾病流行病学教育部重点实验室（北京大学）（北京 100191）
2. 北京大学公共卫生学院流行病与卫生统计学系（北京 100191）
3. 国家儿童医学中心/首都医科大学附属北京儿童医院临床流行病与循证医学中心（北京 100045）
4. 内蒙古自治区药品审评和药物警戒中心（呼和浩特 010010）
5. 北京中医药大学中药学院（北京 102488）
6. 宁波市鄞州区疾病预防控制中心（浙江宁波 315101）
7. 国家药品监督管理局药品评价中心/国家药品监督管理局监管科学创新研究基地（北京 100163）
8. 北京大学医学部药品上市后安全性研究中心（北京 100191）
9. 北京大学第三医院临床流行病学研究中心（北京 100191）
10. 北京大学第三医院眼科（北京 100191）
11. 新疆医科大学中医学院（乌鲁木齐 830017）
12. 新疆石河子大学公共卫生学院（新疆石河子 832000）

【摘要】随着真实世界数据在药品上市后安全性评价、疗效比较研究、药物利用研究和监管决策中的应用不断扩大，多源异构数据的结构差异、术语映射差异、数据质量差异和分析流程差异，成为影响研究可重复性和证据可信度的重要挑战。通用数据模型（CDM）通过统一数据结构、标准术语和分析接口，为跨机构、跨区域和跨国家的真实世界研究提供了基础支撑。近年来，以OHDSI/HADES和FDA Sentinel等平台为代表的工具体系，推动CDM数据分析从传统单库统计建模，逐步发展为标准化、可重复、可诊断、可分布式执行的网络化证据生成模式。本文基于《药物流行病学研究方法学指南（第2版）》关于真实世界数据质量、研究设计、偏倚控制和统计分析计划的要求，系统梳理CDM数据分析的工具及其方法学流程，重点讨论数据质量自动化评价、队列诊断、活性对照新用药者设计、大规模倾向评分、阴性对照与经验校准、目标试验仿真、患者层面预测模型、分布式网络分析和多数据库证据合成等方法学进展。本文旨在为研究者理解和规范应用CDM工具体系、提升药物流行病学真实世界研究的透明度、可重复性和决策价值提供参考。

【关键词】药物流行病学；方法学；指南；通用数据模型；真实世界证据

【中图分类号】R181.3+5　　**【文献标志码】**A

DOI: 10.12173/j.issn.1005-0698.202606128

基金项目：国家自然科学基金面上项目（72474008）；国家自然科学基金国际（地区）合作与交流项目（72361127500）；海南省科学技术厅重点研发专项（ZDYF2024LCLH002）；国家药监局药品评价中心开放课题（CDR2024R01001）

通信作者：孙凤，博士，研究员，博士研究生导师，Email：sunfeng@bjmu.edu.cn
詹思延，博士，教授，博士研究生导师，Email：siyan-zhan@bjmu.edu.cn

Guide on Methodological Standards in Pharmacoepidemiology in China (2nd edition) and their series interpretation (19): tool systems and methodological paradigms for common data model-based data analysis

WU Yunxiao[1,2], NIE Xiaolu[3], ZHENG Yongqi[1,2], WANG Conghui[4], WU Jiarui[5], SUN Yexiang[6], SONG Haibo[7], ZHAN Siyan[1,2,8,9], SUN Feng[1,2,8,10,11,12]

1. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China
2. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China
3. Center for Clinical Epidemiology and Evidence-Based Medicine, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China
4. Inner Mongolia Center for Drug Evaluation and Pharmacovigilance, Hohhot 010010, China
5. School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing 102488, China
6. Yinzhou District Center for Disease Control and Prevention, Ningbo 315101, Zhejiang province, China
7. Drug Evaluation Center, National Medical Products Administration, NMPA/NMPA Center for Innovation and Research in Regulatory Science, Beijing 100163, China
8. Center of Postmarketing Drug Safety Evaluation, Peking University Health Science Center, Beijing 100191, China
9. Clinical Epidemiology Research Center, Peking University Third Hospital, Beijing 100191, China
10. Department of Ophthalmology, Peking University Third Hospital, Beijing 100191, China
11. College of Traditional Chinese Medicine, Xinjiang Medical University, Urumqi 830017, China
12. School of Public Health, Shihezi University, Shihezi 832000, Xinjiang Uygur Autonomous Region, China
Corresponding authors: SUN Feng, Email: sunfeng@bjmu.edu.cn; ZHAN Siyan, Email: siyan-zhan@bjmu.edu.cn

【Abstract】With the expanding use of real-world data in post-marketing drug safety evaluation, comparative effectiveness research, drug utilization studies, and regulatory decision-making, differences in data structures, terminology mappings, data quality, and analytic workflows across heterogeneous data sources have become major challenges affecting the reproducibility of research and the credibility of evidence. The common data model (CDM) provides foundational support for real-world studies across institutions, regions, and countries by harmonizing data structures, standard terminology, and analytic interfaces. In recent years, tool ecosystems represented by OHDSI/HADES and the FDA Sentinel Initiative have promoted CDM-based data analysis from traditional single databases statistical modeling toward a standardized, reproducible, diagnosable, and distributable model of networked evidence generation. Based on the requirements for real-world data quality, study design, bias control, and statistical analysis plans outlined in the *Guidelines for Pharmacoepidemiologic Research Methods* (*2nd Edition*), this article systematically reviews the tools and methodological workflow for CDM-based data analysis. Particular attention is given to methodological advances in automated data quality assessment, cohort diagnostics, active-comparator new-user designs, large-scale propensity score methods, negative controls and empirical calibration, target trial emulation, patient-level prediction models, distributed network analysis, and evidence synthesis across multiple databases. This article aims to provide a reference for researchers to understand and standardize the application of CDM tool ecosystems and to improve the transparency, reproducibility, and decision-making of pharmacoepidemiologic real-world studies.

【Keywords】Pharmacoepidemiology; Methodology; Guideline; Common data model; Real world evidence

随着医疗健康数据资源不断积累，真实世界数据（real world data，RWD）在药物流行病学研究中的作用日益增强，已成为药品上市后安全性评价、疗效比较研究、药物利用研究和监管决策的重要证据来源[1]，常用RWD来源包括电子健康记录、医保数据、疾病登记系统、药品安全监测数据及区域健康信息平台等。然而，传统真实世界研究多基于单一数据库、单一研究问题和单次统计分析开展，研究结果容易受到本地数据结构、编码习惯、变量定义、数据完整性以及研究者分析选择的影响。即使针对相同药物暴露和临床结局，不同数据库之间也可能因数据来源、术语体系、观察期定义和结局捕获能力不同而产生差异，从而影响研究结果的可重复性和可比性[2-3]。

通用数据模型（common data model，CDM）的提出，为多源异构RWD在数据结构、临床语义和分析流程层面的一致化处理提供了重要路径。CDM能够将不同机构、地区和国家来源的数据转换为可重复分析的数据基础，使同一研究问题能够在多个数据库中按照统一队列定义、风险窗口、协变量窗口和统计代码执行[4-5]。近年来，国际药物流行病学和真实世界证据（real world evidence，RWE）研究的发展重点已逐渐从单纯的“CDM建库”转向“CDM方法工具体系”建设，以观察性健康数据科学与信息学协作组织（Observational Health Data Sciences and Informatics，OHDSI）为代表的国际协作网络，旨在通过开放源代码数据分析方案和大规模健康数据库网络生成可靠证据；在此基础上，围绕数据质量诊断、队列定义与表型评价、因果效应估计、患者层面风险预测、安全性信号检测、分布式分析和多数据库证据合成等关键环节，逐渐形成相对完整的工具链和方法学框架[6-8]。

因此，本文基于《药物流行病学研究方法学指南（第2版）》（以下简称“指南第2版”）关于RWD质量、研究设计、偏倚控制和统计分析计划的基本要求，重点解读CDM数据分析相关的新工具体系及其方法学流程。全文首先阐述CDM如何通过数据结构、临床语义和分析流程标准化支撑可重复研究，随后介绍开放源代码体系中的数据质量评价、队列诊断、因果推断和证据合成方法，进一步讨论监管导向分布式网络在药品安全性主动监测中的应用，最后结合现有的平台化工具和目标试验仿真（target trial emulation，TTE）框架，讨论CDM数据分析从“模型自动化”走向“设计结构化”的发展趋势，以期为研究者规范开展CDM驱动的药物流行病学研究、提升RWE的透明度、可重复性和决策价值提供参考。

1 CDM驱动的分析标准化

在药物流行病学研究中，统计分析的可重复性和可比性首先取决于数据结构与临床语义是否一致，而不仅取决于所采用的统计模型。以观察性医疗结果合作项目通用数据模型（Observational Medical Outcomes Partnership Common Data Model，OMOP CDM）为例，它通过统一的数据表结构和标准化临床概念，将研究对象、药物暴露、临床结局、协变量、观察期和随访窗口等关键研究要素转化为可重复定义和提取的分析对象[9]。因此，CDM的意义并非简单地“更换数据库表结构”，而是为后续队列构建、变量定义和统计分析提供一致的数据基础。

仅有结构标准化仍不足以保证分析结果可比。RWD中的诊断、药品、实验室检查、医疗操作等信息，往往依赖本地编码体系和记录习惯；如果不同数据库中相同临床概念不能被映射到统一术语体系，即使采用相同统计模型，也可能分析的并不是完全相同的暴露或结局。OHDSI标准词汇体系是OMOP CDM的核心组成部分，可用于组织和标准化不同临床领域的医学术语，并支持在构建暴露、结局表型及其他研究特征时开展标准化分析[10]。因此，CDM驱动的分析标准化至少包含两个层面：一是表结构和字段层面的结构标准化，二是临床概念和研究变量层面的语义标准化。

在此基础上，CDM进一步推动统计分析流程的标准化。对于同一研究问题，研究者可以预先定义队列进入日期、清洗期、暴露风险窗口、协变量评估窗口、结局判定窗口和删失规则，并将这些设计要素固化为可重复执行的分析程序。美国食品药品监督管理局（Food and Drug Administration，FDA）建立的Sentinel System是较早用于药品上市后主动安全性监测的分布式数据网络，其Sentinel CDM旨在使数据合作方能够在本地数据上快速执行分布式程序[4]；欧盟真实世界数据分析与查询网络（Data Analysis and Real World Interrogation Network，DARWIN EU）也强调通过

将数据标准化为OMOP CDM，使各数据合作方在本地运行统一分析程序，仅共享汇总结果而非患者级数据[5]。这种模式使多中心研究不必依赖患者级数据集中汇聚，而是通过统一数据模型、统一队列定义和统一分析代码，在保护数据隐私的同时提高研究结果的可重复性和可扩展性。

我国药品监管部门对RWD和RWE形成了以数据适用性和研究设计透明度为核心的技术要求。国家药监局药审中心2021年发布的《用于产生真实世界证据的真实世界数据指导原则（试行）》要求从相关性和可靠性两个方面评价数据适用性，并对数据提取、清洗、转换（包括CDM和医学编码）、传输、存储及质量控制等治理环节进行规范；2023年发布的《药物真实世界研究设计与方案框架指导原则（试行）》进一步要求在研究方案中预先明确研究目的、目标人群、暴露与对照、结局、时间窗、潜在偏倚及控制方法、统计分析和敏感性分析等关键要素[11-12]。与Sentinel和DARWIN EU通过特定CDM及分布式程序实现网络层面标准化不同，我国现行指导原则并未限定必须采用某一种特定CDM，而是强调数据是否适用于特定研究问题、数据治理过程是否可追溯，以及研究设计和分析是否透明。

从证据生成流程看，CDM是连接原始RWD与统计分析工具的关键基础（图1）。电子健康记录、医保数据、登记系统和药品安全监测数据等多源数据，需经过结构转换、术语映射和质量评估后，才能支持队列诊断、药物利用分析、因果效应估计、患者层面预测、安全性信号检测和多数据库证据合成。近年来，围绕CDM和分布式RWD分析，形成了多个具有代表性的工具体系和研究网络，包括OHDSI及其健康分析数据生成证据套件（Health Analytics Data-to-Evidence Suite，HADES）、Sentinel、DARWIN EU、美国患者中心临床研究网络（National Patient-Centered Clinical Research Network，PCORnet）、整合生物学与床旁医学信息学平台/共享健康研究信息网络（Informatics for Integrating Biology and the Bedside/Shared Health Research Information Network，i2b2/SHRINE）等。

2 开放源代码体系中的数据诊断、因果推断与证据合成

在众多CDM相关工具体系中，OHDSI/HADES是开放源代码方法体系的代表。HADES的目标是将CDM数据分析过程工具化、可复现和

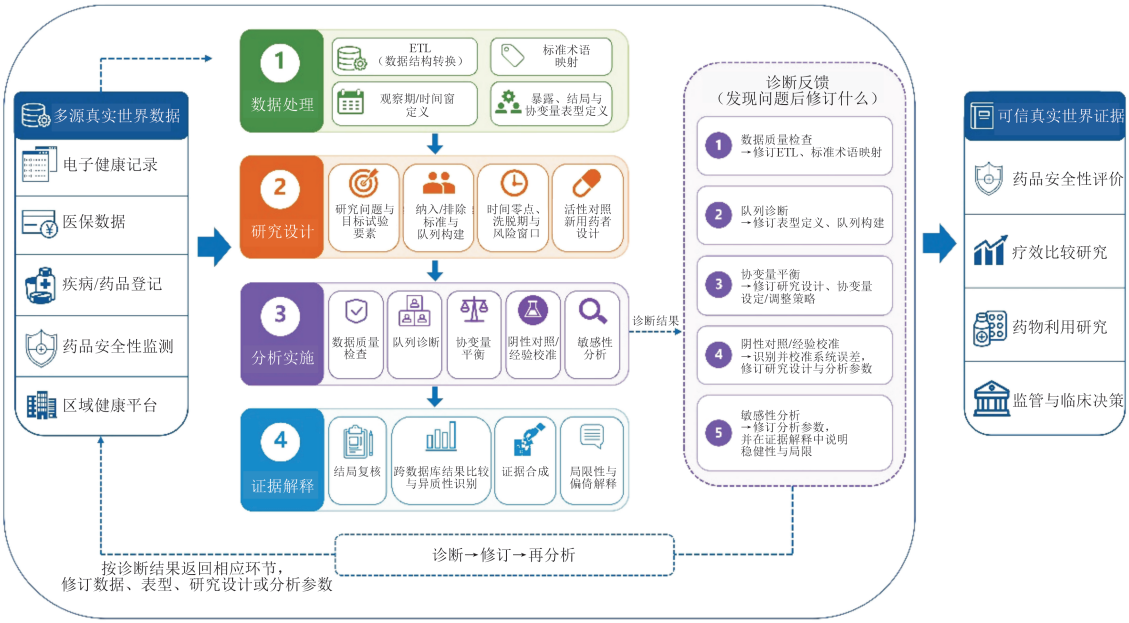


图1 CDM驱动的真实世界证据生成诊断闭环

Figure 1. Diagnostic loop for CDM-driven real world evidence generation

注：CDM. 通用数据模型；RWE. 真实世界证据；OHDSI. 观察性健康数据科学与信息学协作组织；HADES. 健康分析数据生成证据套件；Sentinel. 美国食品药品监督管理局主动监测系统；DARWIN EU. 欧盟真实世界数据分析与查询网络；ACNU. 活性对照新用药者设计；TTE. 目标试验仿真；ETL. 抽取、转换和加载。

可审查化[6]。研究者可通过ATLAS这一基于Web的交互式分析设计工具构建概念集和队列定义，并结合DataQualityDashboard、ACHILLES、CohortDiagnostics、CohortMethod、SelfControlledCaseSeries、PatientLevelPrediction和EvidenceSynthesis等工具，完成从数据质量检查、队列诊断、因果效应估计、系统误差评价到多数据库证据整合的连续分析过程[6]。

数据质量评价是CDM分析的前置环节。虽然CDM能够统一数据结构和术语体系，但数据完成映射后仍可能存在字段缺失、值域异常、时间顺序不合理、概念映射错误或本地记录习惯差异等问题。DataQualityDashboard是OHDSI体系中用于评价OMOP CDM数据质量的代表性工具，其设计基于Kahn等提出的数据质量框架，主要从一致性、完整性和合理性等维度识别潜在质量问题[13]。这一环节的价值在于，研究者不再仅凭经验判断数据库是否适合分析，而是可以用可重复的诊断规则系统检查数据质量，并将诊断结果作为解释研究局限性和跨数据库差异的重要依据。

在数据质量评价之后，队列诊断与表型评价构成CDM研究的第二个关键环节。药物流行病学研究中的目标人群、药物暴露、对照、结局和协变量通常依赖可计算表型定义。可计算表型是指利用诊断代码、诊断术语、用药记录、实验室指标、就诊类型和时间窗等可程序化信息，在数据库中识别具有特定临床特征的人群。由于不同数据源对同一临床概念的记录方式并不完全一致，同一疾病或结局往往存在多个候选识别算法。例如，Yang等[14]基于我国鄞州区域健康信息平台开发并验证了五种自身免疫性疾病的电子健康记录识别算法，发现桥本甲状腺炎和类风湿关节炎可通过相对简单的诊断术语或编码获得较好识别性能，而炎症性肠病和原发免疫性血小板减少症则需要整合门诊、住院、出院记录中的ICD-10编码和中文诊断术语，才能达到可接受的灵敏度和阳性预测值。因此，表型算法并非简单选择一个诊断代码，而是需要结合疾病特点、数据记录方式和研究目的进行开发与验证。在CDM框架下，CohortDiagnostics可进一步生成队列进入事件代码分布、发生率、就诊类型分布及队列进入前后临床事件分布等信息，用于比较不同队列定义在多个数据源中的表现[15]。Rao等[15]利用CohortDiagnostics对系统性红斑狼疮和阿尔茨海默病表型算法进行跨数据源评价，比较不同算法捕获人群的发生率、入组代码和临床特征分布，从而判断表型算法是否符合预期疾病特征。因此，在开放源代码CDM体系中，队列定义不再只是数据提取步骤，而是需要被诊断、比较和报告的方法学对象。研究者除报告最终样本量外，还应说明纳入排除流程、触发入组的概念代码、基线特征、合并用药、结局发生率及其在不同数据库中的差异。

在人群层面的因果效应估计方面，OHDSI/HADES强调以活性对照新用药者设计（Active-Comparator New-User，ACNU）为基础开展比较效应研究。该设计结合了两项互补策略：活性对照部分选择与目标治疗具有相似适应证、处于相近治疗决策节点的替代治疗使用者作为对照，主要提高治疗适应证、疾病严重程度和就医行为的可比性，减少适应证混杂；新用药者部分要求研究对象在预设清洗期内未使用目标药物或对照药物，并从首次用药开始纳入，主要减少现用药者偏倚，同时有利于完整测量治疗前协变量和捕获早期结局。需要注意在使用ACNU设计时，只有将纳入资格判定、治疗分配（即暴露组归属）和随访起点严格对齐于同一时间零点，且仅使用该时点已经获得的信息进行分组，才能避免由时间起点错位造成的永恒时间偏倚[16]。在此设计基础上，CohortMethod等工具可进一步结合大规模倾向评分方法，利用CDM中诊断、用药、操作、检验、合并症等大量基线协变量进行混杂控制。其中，大规模倾向评分并非单一固定算法，而是指在高维医疗数据环境中利用大量诊断、用药、操作、检验和合并症等变量估计治疗分配概率的一类方法。其实现路径可包括高维倾向评分（high-dimensional propensity score，hdPS）、大规模正则化回归以及机器学习模型等。Schneeweiss等[17]提出的hdPS方法可从医疗理赔数据的多个维度中自动筛选经验协变量；Tian等[18]研究通过真实世界和合成数据实验评价了大规模倾向评分方法的表现，为在高维观察性医疗数据中开展倾向评分估计提供了方法学依据。但倾向评分方法的重点不应仅是得到一个校正后的效应值，研究方案和结果至少应说明：①目标估计量及其对应的目标人群，如总体平均治疗效应、已治疗者平均治

疗效应或重叠人群平均治疗效应；②倾向评分模型、协变量选择及其测量时间窗；③具体匹配或加权方法，包括匹配比例、是否放回、卡钳宽度及其尺度，或权重公式及是否稳定化；④共同支持域限制、样本修剪或权重截断阈值，以及由此排除的研究对象；⑤调整前后的倾向评分或权重分布、协变量标准化差异、极端权重和有效样本量；⑥结局模型和方差估计方法。由于不同匹配或加权方法对应的目标人群可能不同，目标估计量应依据临床问题预先设定，并与后续分析方法保持一致。

除传统队列研究外，HADES还包括SelfControlledCaseSeries等自身对照设计工具，可用于短期暴露与急性结局相关的药物安全性研究。自身对照病例系列设计（self-controlled case series，SCCS）使个体作为自身对照，比较同一人在暴露风险期与对照期的事件发生率，从而控制不随时间变化的个体层面混杂因素[19]。这类方法特别适用于疫苗安全性、短期用药不良事件等场景，但标准SCCS的适用依赖若干关键假设：事件发生不应显著改变后续暴露概率，也不应终止或缩短个体的观察过程；暴露日期、事件日期和风险窗口应记录准确；对于非复发事件通常分析首次事件，对于复发事件则需预先说明事件计数方式及事件间独立性假设，并在复发事件存在相关性时采用相应的扩展模型；年龄、季节等随时间变化的混杂因素也应通过分段或参数化方式适当控制。若结局会影响后续治疗、死亡或失访，标准SCCS估计可能产生偏倚，应考虑设置暴露前风险期、采用事件依赖暴露或事件依赖观察期的扩展方法，并开展敏感性分析[20]。

OHDSI/HADES工具中另一个重要方法学特色，是通过阴性对照、阳性对照和经验校准评价研究设计的系统误差。传统观察性研究通常报告点估计、95%置信区间和*P*值，但这些统计量主要反映随机误差，难以充分揭示残余混杂、选择偏倚和错分偏倚等系统误差。经验校准方法利用一组被认为不存在真实因果效应的阴性对照问题估计特定研究设计下的经验零分布，从而校准*P*值或置信区间，使结果解释同时考虑随机误差和系统误差[21-22]。这一思路体现了CDM工具体系的重要转变，即研究者不仅要判断模型结果是否具有统计学意义，还需要评估当前数据、队列定义和分析设计是否足以产生可信证据。

在多数据库网络研究中，效应估计通常还需进一步进入证据合成环节。各数据源可在本地运行相同队列定义和分析程序，分别生成效应估计、协变量平衡、阴性对照、经验校准和敏感性分析等结果；随后可通过固定效应或随机效应模型汇总效应量，并结合数据库来源、患者结构、暴露和结局捕获能力、随访长度及数据质量诊断结果解释异质性。在此背景下，OHDSI发起了大规模跨数据库证据生成与评价（Large-scale Evidence Generation and Evaluation across a Network of Databases，LEGEND）研究倡议。LEGEND并非单一统计模型或软件，而是一套面向成组研究问题的标准化证据生成范式：在分析前预设研究问题和分析方案，在多个异质数据库中使用开放源代码一致执行，利用对照问题和经验校准评价系统误差，并不依据估计结果是否显著进行选择性报告，而是系统公开全部生成的证据[7]。因此，CDM网络研究中的证据合成并非简单“合并数据”或“合并效应值”，而是需要同时整合效应估计、设计诊断、系统误差评价和数据库间异质性信息。

综上，OHDSI/HADES代表了CDM数据分析从单一模型应用向标准化证据生成流程的转变。其方法学价值不仅在于提供开放源代码工具，也在于将数据质量评价、队列诊断、因果效应估计、系统误差评价和多数据库证据合成纳入同一可复核流程。这一点也为后续监管导向分布式网络的发展提供了重要方法学参照。

3 分布式监管网络中的药品安全性监测

与OHDSI/HADES更强调开放源代码的方法学流程不同，Sentinel和DARWIN EU代表了监管问题驱动的分布式RWE生成模式。本节主要围绕监管网络中的四类分析任务展开：第一，利用模块化程序开展快速查询和描述性分析；第二，针对预设安全性问题开展观察值/期望值（observed-to-expected，O/E）分析和序贯监测；第三，针对未知潜在风险利用树状扫描统计量（tree-based scan statistic，TreeScan）等方法进行信号发现；第四，通过DARWIN EU等跨国网络支持监管问题的快速研究和证据解释。

Sentinel的一个重要方法学特点是将常见监管

分析任务程序化。对于初步安全性问题，可先通过模块化查询程序快速描述目标药品使用人群、暴露规模、合并用药、基线疾病和结局发生情况；当需要进一步评估暴露与结局关联时，则可根据问题特点选择队列设计、SCCS、O/E 分析或序贯监测等方法。Gagne 等[23]比较了 Sentinel 模块化程序与传统研究程序在重现药物安全性研究中的表现，提示参数化分析工具有助于在分布式数据库环境中提高研究执行效率和结果一致性。这种模式的价值在于将部分常规药物流行病学分析转化为可配置、可重复运行的程序，同时保留针对具体问题制定研究方案和解释结果的必要空间。

在疫苗和其他需持续监测的药品安全性评价中，序贯分析是分布式监管网络的重要方法。传统一次性分析通常在数据收集完成后进行，而序贯监测允许在数据持续更新过程中反复比较观察事件数与期望事件数，以便更早发现潜在风险。最大序贯概率比检验是药物和疫苗安全性监测中常用的序贯方法，可用于在连续监测场景中控制反复检验带来的错误发现问题[24]。Yih 等[25]在 Sentinel System 中开展了关于流感疫苗的准实时安全性主动监测，表明 Sentinel 可用于建立持续更新的序贯安全性监测系统，以及时评估流感疫苗及其他医学应急产品的潜在风险。这类方法尤其适用于风险窗口相对明确、背景发生率可估计且公共卫生意义较高的安全性问题。

对于未知的潜在安全性信号，TreeScan 提供了另一类信号发现路径。与针对单一预设结局的队列研究不同，TreeScan 可利用诊断代码或医学概念的层级结构，在大量候选结局中搜索异常聚集信号[26]。Yih 等[27]使用自我对照树状-时间扫描统计量评估四价 HPV 疫苗安全性，指出该方法可在不预先限定具体健康结局和风险窗口的情况下，对多种潜在结局进行信号检测。这类方法能够扩大安全性监测的覆盖范围，但其结果更适合作为信号生成和优先级排序，不能直接等同于因果确认；阳性信号仍需结合临床合理性、病历复核、既往证据和验证性研究进一步判断。

欧洲监管体系下的 DARWIN EU 则代表了以监管需求驱动的跨国 RWE 网络。欧洲监管机构正在推动 RWE 在药品监管中的应用，使其覆盖上市前后更广泛的决策场景[28]。与 Sentinel 主要以 FDA 主动上市后安全性监测为核心不同，DARWIN EU 更强调研究问题由监管需求触发、协议和分析流程预先设定、数据在本地运行并形成供监管机构解释的研究报告。这种模式有助于提高监管问题响应速度，但也要求对跨国家、跨医疗系统的数据差异进行充分评估。除监管网络外，PCORnet 和 i2b2/SHRINE 代表了另一类偏向临床研究和队列发现的分布式数据基础设施。PCORnet 主要服务于患者中心结局和比较效果研究，SHRINE 则是在多个 i2b2 站点之间开展联邦查询的工具，可帮助研究者在不集中患者级数据的情况下估算符合特定条件的人群规模并评估研究可行性。相较于 Sentinel 和 DARWIN EU，此类网络更适合研究前期的队列发现和可行性评估，不能直接替代完整的因果推断研究。

总体来看，分布式监管网络的优势在于能够在不集中共享患者级数据的前提下，利用多数据源开展主动安全性监测和快速证据生成；其局限则在于不同数据源之间的编码体系、随访完整性、结局记录及时性和医疗实践差异仍可能影响结果解释。因此，这类网络并不能替代传统药物流行病学判断，而是将监管问题、分布式数据、标准化分析和证据解释整合到更快速、更可审查的工作流程中。

4 平台化分析工具与目标试验仿真的方法学融合

除开放源代码体系和监管导向分布式网络外，平台化 RWE 工具代表了另一种研究实施模式。其重点不在于建立新的 CDM，而是在标准化或预处理后的 RWD 基础上，将队列构建、变量定义、混杂控制、结局分析、报告生成和审计追踪等步骤整合到统一操作环境中[29-30]。这类平台进一步承接了 CDM 的标准化数据基础，将研究设计、队列定义、分析参数和结果输出转化为可执行、可追踪的研究实施流程。与研究者完全从头编写代码相比，平台化工具的优势在于提高研究执行效率，并通过预设参数、自动化文档和审计轨迹增强分析过程的可追溯性。例如，一项旨在通过实证复现评价医疗数据库研究报告透明度和可重复性的研究计划使用 Aetion Evidence Platform 及其他统计软件复现了 150 项已发表数据库研究，该平台可自动记录时间锚点、代码算法等研究实施细节和

审计轨迹，结果显示原始效应量与复现效应量呈较强相关，提示透明化流程和结构化工具有助于提升RWE研究的可复现性[29]。TriNetX则是面向生物医学和临床研究的全球联邦式RWD分析平台，可在不集中汇聚患者级数据的前提下，支持跨医疗机构的队列识别、特征描述和RWD分析[30]。

但平台化工具的价值不应被理解为“点击式生成证据”。真实世界研究能否产生可信结论，关键仍在于研究问题是否清晰、暴露和结局能否准确测量、混杂因素是否充分捕获以及随访起点和风险窗口是否设置合理。因此，平台化RWE工具的真正方法学价值，不是替代研究设计本身，而是帮助研究者在标准化数据基础上，将预先定义的研究问题、队列规则、变量算法、分析参数和结果输出转化为可执行、可追踪、可复核的研究流程。沿着这一逻辑，TTE可作为CDM和平台化RWE分析中连接“标准化数据”与“结构化研究设计”的重要框架。Hernán和Robins提出，当随机对照试验不可行时，基于观察性数据的因果分析应首先明确一个理想随机对照试验的核心要素，包括目标人群、治疗策略、分配方式、随访起点、结局、因果对比和分析计划，再将这些要素映射到可获得的RWD中[31]。对于CDM数据而言，这一框架的价值在于推动研究者在运行模型之前，先将入排标准、暴露定义、对照选择、风险窗口、协变量窗口和删失规则转化为可执行、可复核的队列和变量定义，从而减少随访起点、纳入标准和治疗分配时间不一致所导致的永恒时间偏倚等问题[32]。平台化工具与TTE的结合，代表了RWE分析从“分析自动化”进一步走向“设计结构化”。CDM或平台化工具提供标准化变量来源、队列构建界面、参数留痕和可重复执行环境，TTE则提供因果问题的设计蓝图；二者结合后，研究者可以更清楚地说明目标人群如何定义、治疗策略如何比较、随访从何时开始、结局如何测量以及哪些分析用于检验假设稳健性。对于治疗策略需要在随访早期进一步确定或涉及治疗改变和依从性问题的研究，可在目标试验框架下结合克隆-删失-加权、逆概率治疗加权或逆概率删失加权等方法，但这些方法应服务于预先定义的目标试验方案，而不应取代研究设计本身。近年来，TARGET报告规范进一步提出了针对显式模拟目标试验的观察性研究报告清单，强调应说明因果问题、目标试验方案、目标试验与观察性数据之间的映射关系、估计量及敏感性分析等内容[33]。这提示未来平台化RWE工具不仅应提高分析效率，更应支持研究方案预注册、关键参数留痕、可重复代码导出和透明报告。为便于对前述工具体系和方法学流程进行整体比较，表1总结了主要工具体系的定位、代表性功能、适用场景和方法学注意事项。

表1 主要CDM相关工具、网络与平台的层级定位及方法学特征

Table 1. Positioning and methodological characteristics of major CDM-related tools，networks，and platforms

CDM相关的工具体系或平台	类型与层级定位	代表性功能或组成	适用药物流行病学场景	方法学注意事项
OHDSI /OMOP CDM /HADES	开放科学协作网络、通用数据模型和开放源代码分析工具体系	OMOP CDM、标准词汇体系、ATLAS、DataQualityDashboard、CohortDiagnostics、CohortMethod、SelfControlledCaseSeries、EmpiricalCalibration、EvidenceSynthesis等	药物利用研究、疗效比较研究、安全性评价、患者层面风险预测和多数据库证据合成	依赖高质量ETL、术语映射和表型算法验证；需报告数据质量评价、队列诊断、协变量平衡、经验校准和多数据库异质性
Sentinel System/ Sentinel CDM	FDA监管导向的主动安全监测系统及其数据模型	Sentinel CDM、模块化查询程序、O/E分析、序贯监测、TreeScan、分布式本地运行等	药品和疫苗上市后安全性主动监测、监管问题快速响应、安全性信号评估	需关注结局捕获延迟、背景发生率估计、风险窗口设定、数据刷新频率和信号验证

续表1

CDM相关的工具体系或平台	类型与层级定位	代表性功能或组成	适用药物流行病学场景	方法学注意事项
DARWIN EU	欧洲监管导向的分布式真实世界数据网络，主要基于OMOP CDM支持跨数据源研究	标准化研究方案、本地运行统一分析程序、汇总结果共享、跨国家多数据源研究报告	药品全生命周期监管问题、上市前后证据补充、药品安全性和有效性相关RWE生成	需关注不同国家和医疗系统间的数据结构、编码习惯、医疗实践和结局捕获差异；监管解释需结合数据源适用性和异质性
PCORnet、i2b2/SHRINE	临床研究网络与联邦查询基础设施	队列构建、病例数估算、多中心可行性评估、患者中心结局研究支持、联邦查询	实效性研究准备、患者中心结局研究、多中心队列构建和研究可行性评估	更适合前期队列构建和可行性判断，不能直接替代完整因果推断研究；后续仍需明确研究方案、变量定义和统计分析计划
Aetion、TriNetX等平台化RWE工具	平台化研究实施与分析工具	界面化队列构建、变量定义、基线特征描述、倾向评分分析、结局模型、自动化报告、审计轨迹或联邦式分析	产业界RWE研究、监管沟通、卫生技术评估、市场准入研究和快速探索性分析	不能替代研究设计；需透明报告数据来源、平台版本、变量定义、模型参数和分析流程，避免“点击式生成证据”；可融合TTE方法作为因果研究的结构化设计框架

注：CDM.通用数据模型；OHDSI.观察性健康数据科学与信息学协作组织；OMOP.观察性医疗结果合作项目通用数据模型；HADES.健康分析数据生成证据套件；ETL.抽取、转换和加载；O/E.观察值/期望值；FDA.美国食品药品监督管理局；TreeScan.树状扫描统计量；DARWIN EU.欧盟真实世界数据分析与查询网络；PCORnet.美国患者中心临床研究网络；i2b2.整合生物学与床旁医学信息学平台；SHRINE.共享健康研究信息网络；TTE.目标试验仿真。

5 结语

本文围绕CDM数据分析的工具体系和方法学流程，梳理了开放源代码体系、监管导向分布式网络、平台化RWE工具以及TTE等近年来的重要进展。总体来看，CDM及相关工具的发展正在改变药物流行病学研究的实施方式，使研究从传统的单库分析逐步转向以标准化队列定义、可重复分析代码、诊断性质量评价和多数据库证据整合为特征的证据生成流程。然而，工具体系的成熟并不意味着研究证据天然可靠。CDM转换后的数据仍可能存在源数据缺失、术语映射偏差、结局捕获不足、暴露时间记录不准确和跨数据库异质性等问题；即使分析流程高度标准化，若研究问题、目标人群、变量定义和因果对比设定不清，仍可能产生偏倚结果。因此，未来CDM数据研究应更加重视“适用性先于分析”的原则，即在模型运行前系统评估数据是否足以回答特定研究问题，并将数据质量评价、队列诊断、设计诊断和敏感性分析作为结果解释的必要组成部分。

在我国，CDM及其工具体系具有重要发展潜力。电子病历、医保数据、疾病登记、区域健康信息平台和药品安全监测数据等资源日益丰富，为药品上市后安全性评价、疗效比较研究、药物利用研究和卫生决策提供了广泛数据基础，但数据获取、共享和质量仍是持续挑战[34-35]。同时，RWE在我国药品监管决策中的应用已得到更多关注，但未来仍需进一步完善数据质量、研究设计和监管接受度等关键环节[36]。因此，我国CDM建设不应停留在单个医院或单一区域的数据转换，而应逐步走向跨机构、跨区域、可审查的分布式研究网络，并且至少需要在以下三个方面持续深化。首先，应加强本土术语和数据结构的适配，包括诊断编码、医保目录、药品通用名与商品名、检验单位、手术操作、国产药品以及中成药等本土化语义元素的标准化映射，避免因概念映射不充分而影响暴露、结局和协变量定义。其次，应建设可共享的队列定义、表型算法和数据质量评价规则库，优先围绕药品安全性主动监测、慢病药物疗效比较、罕见病、儿童用药、中成药上市后评价等高需求场景形成可复用模板。另外，应在隐私保护和数据安全前提下发展分布式分析模式，使各数据持有机构能够在本地运行统一程序并共享汇总结果，从而在不集中转移患者级数据的情况下开展多中心研究。

此外，CDM工具体系的推广还依赖跨学科人才和协作机制。CDM数据分析不仅是数据库建设或统计建模问题，还涉及临床医学、药学、流行病学、生物统计、医学信息学、数据治理和监管科学等多学科合作。未来应加强研究者对CDM、TTE、阴性对照、经验校准、序贯监测和分布式分析等方法的系统培训，推动研究方案、队列定义、分析代码和诊断结果的透明报告。一项针对我国生物医学专业人员的调查显示，受访者总体

上认可CDM在中国区域数据库中的应用价值，但也提示需要加强CDM相关教育培训并建立明确的数据共享规则[37]。这说明，方法工具的普及不仅需要技术平台，也需要共同语言、治理规则和长期协作网络。

综上，CDM驱动的药物流行病学研究正在从数据标准化走向证据生成流程标准化。针对我国现状，其关键任务不是简单移植国际工具，而是在借鉴OHDSI/HADES、Sentinel、DARWIN EU等经验的基础上，结合本土数据环境和监管需求，建立适合我国医疗体系的CDM数据质量评价、队列诊断、分布式分析和透明报告规范。唯有在严谨研究设计和持续质量控制的基础上应用这些工具，才能真正提升RWE的可信度、可重复性和决策价值，为药品上市后安全性监测、疗效比较研究、合理用药和监管科学发展提供更可靠的方法学支撑。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 Hennessy S, Atsuta Y, Hill S, et al. Real-world data and real-world evidence in regulatory decision making: report summary from the Council for International Organizations of Medical Sciences (CIOMS) Working Group XIII[J]. Pharmacoepidemiol Drug Saf, 2025, 34(3): e70117. DOI: 10.1002/pds.70117.

2 Finster M, Wenzel M, Taghizadeh E. Common data models and data standards for tabular health data: a systematic review[J]. BMC Med Inform Decis Mak, 2025, 25(1): 422. DOI: 10.1186/s12911-025-03267-2.

3 Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases[J]. J Am Med Inform Assoc, 2015, 22(3): 553-564. DOI: 10.1093/jamia/ocu023.

4 Sentinel Initiative. Sentinel Common Data Model[EB/OL] [2026-06-03]. https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model.

5 DARWIN EU. Data Interoperability[EB/OL] [2026-06-03]. https://darwin-eu.org/index.php/methods/improving-interoperability.

6 Schuemie M, Reps J, Black A, et al. Health-analytics data to evidence suite (HADES): open-source software for observational research[J]. Stud Health Technol Inform, 2024, 310: 966-970. DOI: 10.3233/shti231108.

7 Schuemie MJ, Ryan PB, Pratt N, et al. Principles of large-scale evidence generation and evaluation across a network of databases (LEGEND)[J]. J Am Med Inform Assoc, 2020, 27(8): 1331-1337. DOI: 10.1093/jamia/ocaa103.

8 Hripcsak G, Duke JD, Shah NH, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers[J]. Stud Health Technol Inform, 2015, 216: 574-578. https://pubmed.ncbi.nlm.nih.gov/26262116/.

9 the OHDSI CDM Working Group. OMOP Common Data Model[EB/OL]. [2026-06-04]. https://ohdsi.github.io/CommonDataModel/.

10 Reich C, Ostropolets A, Ryan P, et al. OHDSI standardized vocabularies—a large-scale centralized reference ontology for international data harmonization[J]. J Am Med Inform Assoc, 2024, 31(3): 583-590. DOI: 10.1093/jamia/ocad247.

11 国家药品监督管理局. 用于产生真实世界证据的真实世界数据指导原则(试行)[EB/OL] (2021-04-13)[2026-07-17]. https://www.cde.org.cn/main/news/viewInfoCommon/2a1c437ed54e7b838a7e86f4ac21c539.

12 国家药品监督管理局. 药物真实世界研究设计与方案框架指导原则(试行)[EB/OL] (2023-02-17)[2026-07-17]. https://www.nmpa.gov.cn/xxgk/ggtg/ypggtg/ypqtggtg/20230217143016157.html?type=pc&m=.

13 Blacketer C, Defalco FJ, Ryan PB, et al. Increasing trust in real-world evidence through evaluation of observational data quality[J]. J Am Med Inform Assoc, 2021, 28(10): 2251-2257. DOI: 10.1093/jamia/ocab132.

14 Yang J, Wu Y, Guo J, et al. Development and validation of identification algorithms for five autoimmune diseases using electronic health records: a retrospective cohort study in China[J]. Front Immunol, 2025, 16: 1541203. DOI: 10.3389/fimmu.2025.1541203.

15 Rao GA, Shoaibi A, Makadia R, et al. CohortDiagnostics: Phenotype evaluation across a network of observational data sources using population-level characterization[J]. PLoS One, 2025, 20(1): e0310634. DOI: 10.1371/journal.pone.0310634.

16 Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application[J]. Curr Epidemiol Rep, 2015, 2(4): 221-228. DOI: 10.1007/s40471-015-0053-5.

17 Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data[J]. Epidemiology, 2009, 20(4): 512-522. DOI: 10.1097/EDE.0b013e3181a663cc.

18 Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments[J]. Int J Epidemiol, 2018, 47(6): 2005-2014. DOI: 10.1093/ije/dyy120.

19 Petersen I, Douglas I, Whitaker H. Self controlled case series methods: an alternative to standard epidemiological study designs[J]. BMJ, 2016, 354: i4515. DOI: 10.1136/bmj.i4515.

20 Nie X, Xu L, Bai Y, et al. Self-controlled case series design in vaccine safety: a systematic review[J]. Expert Rev Vaccines, 2022, 21(3): 313-324. DOI: 10.1080/14760584.2022.2020108.

21 Schuemie MJ, Hripcsak G, Ryan PB, et al. Robust empirical calibration of p-values using observational data[J]. Stat Med, 2016, 35(22): 3883-3888. DOI: 10.1002/sim.6977.

22 Schuemie MJ, Hripcsak G, Ryan PB, et al. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data[J]. Proc Natl Acad Sci U S A, 2018, 115(11): 2571-2577. DOI: 10.1073/pnas.1708282114.

23 Gagne JJ, Han X, Hennessy S, et al. Successful comparison of US Food and Drug Administration Sentinel Analysis Tools to traditional approaches in quantifying a known drug-adverse event association[J]. Clin Pharmacol Ther, 2016, 100(5): 558-564. DOI: 10.1002/cpt.429.

24 Kulldorff M, Davis RL, Kolczak M, et al. A maximized sequential probability ratio test for drug and vaccine safety surveillance[J]. Sequential Analysis, 2011, 30(1): 58-78. DOI: 10.1080/07474946.2011.539924.

25 Yih WK, Kulldorff M, Sandhu SK, et al. Prospective influenza vaccine safety surveillance using fresh data in the Sentinel System[J]. Pharmacoepidemiol Drug Saf, 2016, 25(5): 481-492. DOI: 10.1002/pds.3908.

26 孙一鑫, 王淼, 杨明芳, 等. 树状扫描统计量用于药品和疫苗安全性监测的综述[J]. 中华流行病学杂志, 2021, 42(7): 1286-1291. [Sun YX, Wang M, Yang MF, et al. Review on tree-based scan statistic in drug and vaccine safety monitoring[J]. Chinese Journal of Epidemiology, 2021, 42(7): 1286-1291]. DOI: 10.3760/cma.j.cn112338-20201103-01297.

27 Yih WK, Maro JC, Nguyen M, et al. Assessment of quadrivalent human papillomavirus vaccine safety using the self-controlled tree-temporal scan statistic signal-detection method in the Sentinel System[J]. Am J Epidemiol, 2018, 187(6): 1269-1276. DOI: 10.1093/aje/kwy023.

28 Arlett P, Kjaer J, Broich K, et al. Real-world evidence in EU medicines regulation: enabling use and establishing value[J]. Clin Pharmacol Ther, 2022, 111(1): 21-23. DOI: 10.1002/cpt.2479.

29 Wang SV, Sreedhara SK, Schneeweiss S, et al. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions[J]. Nat Commun, 2022, 13(1): 5126. DOI: 10.1038/s41467-022-32310-3.

30 Palchuk MB, London JW, Perez-Rey D, et al. A global federated real-world data and analytics platform for research[J]. JAMIA Open, 2023, 6(2): ooad035. DOI: 10.1093/jamiaopen/ooad035.

31 Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available[J]. Am J Epidemiol, 2016, 183(8): 758-764. DOI: 10.1093/aje/kwv254.

32 Hernán MA, Sauer BC, Hernández-Díaz S, et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses[J]. J Clin Epidemiol, 2016, 79: 70-75. DOI: 10.1016/j.jclinepi.2016.04.014.

33 Cashin AG, Hansford HJ, Hernán MA, et al. Transparent reporting of observational studies emulating a target trial—the TARGET Statement[J]. JAMA, 2025, 334(12): 1084-1093. DOI: 10.1001/jama.2025.13350.

34 Sun F, Bedenkov A, Liu BC, et al. Maximizing the value of real-world data and real-world evidence to accelerate healthcare transformation in China: summary of external advisory committee meetings[J]. Pharmaceut Med, 2024, 38(3): 157-166. DOI: 10.1007/s40290-024-00520-3.

35 Xie J, Wu EQ, Wang S, et al. Real-world data for healthcare research in China: call for actions[J]. Value Health Reg Issues, 2022, 27: 72-81. DOI: 10.1016/j.vhri.2021.05.002.

36 Xu J, Wu W, Zhang X, et al. The use of real-world evidence for regulatory decisions in China[J]. Clin Pharmacol Ther, 2024, 116(1): 82-95. DOI: 10.1002/cpt.3257.

37 Yu Y, Zheng Y, Zhang M, et al. Attitudes toward common data models among Chinese biomedical professionals: cross-sectional survey[J]. JMIR Med Inform, 2025, 13: e77603. DOI: 10.2196/77603.

收稿日期：2026 年 06 月 25 日 修回日期：2026 年 07 月 20 日
本文编辑：桂裕亮 李绪辉