多中心回顾性电子病历数据使用全流程操作规程Δ

钟　雪[1*]，钱东方[2]，张子轩[2]，谭斯元[2]，刘建峰[2]，崔学艳[3]，聂瑞芳[4]，李　婷[5]，王　倩[6]，郭　其[7]，刘秋爽[8]，梁　艳[9]，黄　琳[1]，李理总[1]，封宇飞[1#]（1.北京大学人民医院药学部，北京　100044；2.北京北方医药健康经济研究中心，北京　100021；3.山东省千佛山医院药学部，济南　250014；4.山东省立医院药学部，济南　250021；5.北京医院药学部，北京　100730；6.济南市中心医院药学部，济南　250013；7.滨州医学院烟台附属医院药学部，山东 烟台　256699；8.哈尔滨医科大学附属第二医院药学部，哈尔滨　150086；9.解放军总医院医疗保障中心药剂科，北京　100853）

中图分类号　R951　　**文献标志码**　A　　**文章编号**　1001-0408（2022）19-2314-07
DOI　10.6039/j.issn.1001-0408.2022.19.03

摘　要　随着医疗信息化的逐步完善与医疗健康大数据的蓬勃发展，真实世界研究的探索和实践日益成熟，真实世界数据已成为药品上市后再评价的重要证据来源。电子病历数据作为重要的高质量真实世界医疗数据，更是开展药品上市后再评价研究不可或缺的数据源。现有真实世界研究的指南和规范多从前瞻性研究角度设计，并未提出回顾性研究实施中的具体措施和方法，尤其是针对使用常规收集的电子病历数据技术层面的操作建议。本文结合现有指南和规范制订的操作流程框架，创新性地添加了针对回顾性电子病历数据的数据验证、数据合库、数据核查及贯穿始终的质量控制和数据管理与储存等操作规程，并以药品上市后安全性评价为例，描述利用电子病历数据开展多中心回顾性真实世界研究涉及的数据分析方法及要点，最终建立了适用于多中心回顾性电子病历数据使用的全流程操作规程。

关键词　真实世界数据；电子病历；回顾性研究；数据治理；多中心研究；操作规程

The full process operating procedure for the using of multi-center retrospective electronic medical record data

ZHONG Xue[1], QIAN Dongfang[2], ZHANG Zixuan[2], TAN Siyuan[2], LIU Jianfeng[2], CUI Xueyan[3], NIE Ruifang[4], LI Ting[5], WANG Qian[6], GUO Qi[7], LIU Qiushuang[8], LIANG Yan[9], HUANG Lin[1], LI Lizong[1], FENG Yufei[1]（1. Dept. of Pharmacy, Peking University People's Hospital, Beijing 100044, China; 2. Beijing North Medical Health Economic Research Center, Beijing 100021, China; 3. Dept. of Pharmacy, Shandong Provincial Qianfoshan Hospital, Jinan 250014, China; 4. Dept. of Pharmacy, Shandong Provincial Hospital, Jinan 250021, China; 5. Dept. of Pharmacy, Beijing Hospital, Beijing 100730, China; 6. Dept. of Pharmacy, Jinan Central Hospital, Jinan 250013, China; 7. Dept. of Pharmacy, Yantai Affiliated Hospital of Binzhou Medical University, Shandong Yantai 256699, China; 8. Dept. of Pharmacy, The 2nd Affiliated Hospital of Harbin Medical University, Harbin 150086, China; 9. Dept. of Pharmacy, Medical Supplies Center of PLA General Hospital, Beijing 100853, China）

ABSTRACT　With the gradual improvement of medical informatization and the vigorous development of medical and health big data, the exploration and practice of real-world research are becoming more and more mature, and real-world data have become an important source of evidence for post marketing re-evaluation of drugs. As an important high-quality real-world medical data, electronic medical record data is an indispensable data source for post marketing re-evaluation of drugs. Most of the existing guidelines and norms of real-world research are designed from the perspective of prospective research, and do not propose specific measures and methods in the implementation of retrospective research, especially for the operation suggestions on the technical level of using conventionally collected electronic medical record data. In combination with the operational process framework formulated by the existing guidelines and norms, this paper creatively adds the operating procedures for data validation, data integration, data verification, and throughout quality control, data management and storage of retrospective

Δ **基金项目**　中国药品监督管理研究会立项课题（No.药监研〔2021〕043号）

＊**第一作者**　主管药师，博士。研究方向：药品真实世界研究、免疫药物机制。电话：010-88325750。E-mail：bhdxzhongxue@126.com

\# **通信作者**　主任药师。研究方向：医院药学、药物经济学。E-mail：fengyufei@126.com

electronic medical record data, and describes the data analysis methods and key points involved in carrying out multi-center retrospective real-world research using electronic medical record data, taking the post marketing safety research of drugs as an example. Finally, the full process operation procedure applicable to the use of multi-center retrospective electronic medical record data is established.

KEYWORDS real-world data; electronic medical record data; retrospective research; data curation; multi-center research; operation procedure

真实世界数据来源于患者个人诊疗记录等多种途径，已成为上市后药品安全性监测和评价的重要数据来源。区别于基于研究目的主动收集的健康医疗数据，来源于常规工作中收集的医疗卫生数据虽然具有数据量大、人群覆盖广等优势，但也存在数据缺失、标准化程度不佳等局限[1]。要将真实世界数据转化为真实世界证据以支持临床决策，需要围绕具体研究问题或目的，构建研究型数据处理体系，采用合适的研究设计和统计学方法回答设定好的问题[2]。在将真实世界证据用于为相关领域提供证据的过程中，需要从数据的完整性、一致性、合理性等多角度考虑，从而设计完善的真实世界数据使用操作规范。

2021年4月，国家药品监督管理局药品审评中心发布的《用于产生真实世界证据的真实世界数据指导原则（试行）》对真实世界数据的来源、评价、治理、标准、安全合规、质量保障、适用性等方面给出了具体的要求和建议[3]。2022年7月，国家药品监督管理局药品审评中心发布的《药物真实世界研究设计与方案框架指导原则（征求意见稿）》，对药物研发中真实世界研究的设计及方案制定提出了更有针对性的技术要求[4]。我国流行病学与卫生统计学专家组织翻译的《使用常规收集医疗卫生数据开展观察性研究的报告规范》规范了真实世界研究报告的内容清单，并列举了实例以辅助理解[5-6]。中国真实世界数据与研究联盟于2019年发布了真实世界数据与研究的技术规范，侧重于体系建设与步骤细分[7]。然而，国内现有的指南和规范并未提出研究实施中的具体措施和方法，尤其是针对常规收集的医疗卫生数据在技术层面的操作建议，如数据提取标准、数据结构标准、针对回顾性数据研究设计的相关考虑等。

本研究按照多中心回顾性真实世界研究的流程展开，参考《用于产生真实世界证据的真实世界数据指导原则（试行）》[3]对多中心回顾性电子病历数据使用步骤进行划分（图1），以国内现有的指南和规范提出的标准真实世界数据使用步骤为框架，有针对性地把控和补充真实世界研究各个环节的措施和方法，建立了多中心回顾性电子病历数据使用的全流程操作规范。本文旨在填补现有指南和规范在临床与技术沟通、方案和数据转化、多中心间实时协调等实践层面的空缺，为利用真实世界数据开展药品上市后风险评估过程中的各个环节提供实践指导。

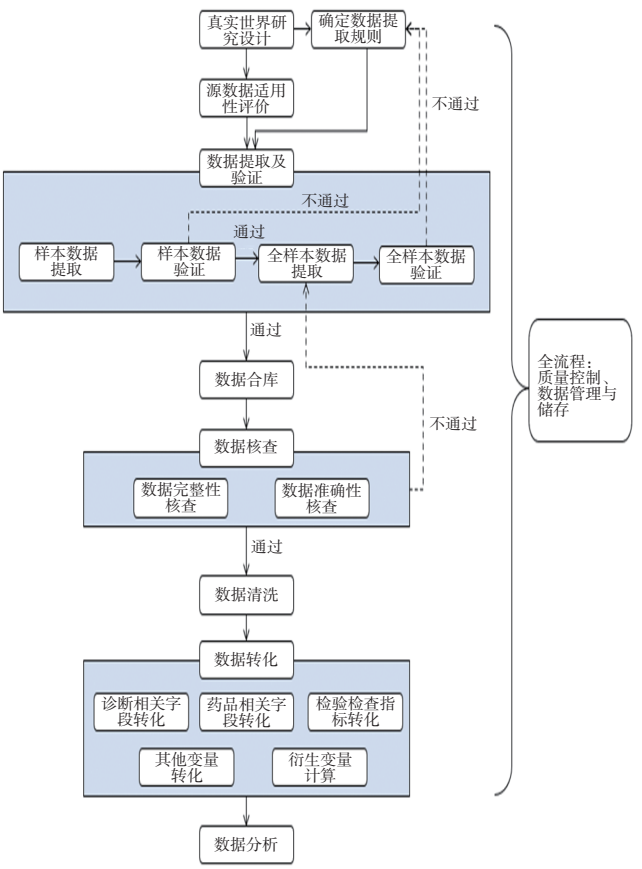

图1 多中心回顾性电子病历数据使用流程图

1 真实世界研究设计

与传统的临床试验相同，真实世界研究开展之初，首先要进行研究方案的确定。明确研究问题后，根据研究目的设计研究内容，特别是纳排标准和研究相关变量（如性别、年龄、用药信息、检验检查信息等），从而确定需提取的字段，设计数据提取规则。当以某种疾病确定研究人群时，疾病的定义建议明确为通识说法，可使用该疾病的国际疾病分类编码（international classification of diseases，ICD）进行描述，并充分考虑纳入编码的准确性和完整性；相关变量的定义建议具体至每一子项，并落实数据提取要求。在设计研究方案时，要结合研究目的与数据库可涵盖的变量，充分了解研究相关的疾病诊断、治疗模式、用药模式，根据临床实际情况选择拟纳入分析的具体字段，如检查子项、合并诊断、合并用药等详细信息，并充分考虑以上信息设计数据提取方案。同时，要明确不同数据库的结构及链接方式，明确所需变

量来源，还应注意设计提取用于链接不同数据表间信息的关键字段。此外，普适性的真实世界研究设计方法在现有指南和规范中已有描述[1,6,8]，本文不再赘述。

2 源数据适用性评价

在初步确定真实世界研究中心后，需对源数据的适用性进行评价。应从可及性、伦理合规、代表性等维度，对源数据进行初步评价和选择。源数据的适用性评价受数据源类型[如医院信息系统数据（与本文关注的电子病历数据属于同类）、患者自报告结局、医保支付数据等]和数据性质（如回顾性或前瞻性）影响较大，较少受真实世界研究目的影响。多中心回顾性电子病历数据适用性评价量表见表1。由于多中心医疗机构常规诊疗数据存在长期随访信息缺失、各中心数据标准不统一等局限性，其他既有健康医疗数据也存在各自的优势和局限性[9]，因此在选择研究中心时，需首要考虑样本量。在条件允许的情况下，应充分考虑地域分布问题，尽可能广泛地纳入具有区域代表性的研究中心。研究参与人员方面，除需具有专业知识背景外，还需了解真实世界研究的方法、有真实世界数据提取与分析经验、熟悉本院数据提取方法及流程，以便在数据处理及分析过程中对于一些本院独有的情况给予合理的解释。经源数据适用性评价无法满足研究需求的中心（如评价证实该中心无法提取或无相关数据），应考虑用其他数据源代替，必要时及时更换研究中心。

表1 多中心回顾性电子病历数据的源数据适用性评价量表

评价对象	评价内容	评价标准
机构	可及性	机构数据是否可被获取
	伦理合规	是否能通过院内伦理委员会审核
	代表性	是否具有区域或学科代表性
	数据状态	是否支持完成项目；立项文件研究方案中是否包括数据提取方法
	关键变量完整性	是否包含临床结局和暴露变量；数据实际情况与拟提取字段表匹配程度
	样本量	是否可以达到目标样本总量
	源数据活动状态	是否可以使用备份数据库提取数据；提取数据是否会影响该中心系统正常使用
人员	研究者	资质背景是否符合研究要求
	信息科人员	是否熟悉回顾性研究数据的提取要求
		对提取数据中异常情况是否能给予合理解释
	联系人	对于本院数据及相关软件、系统的熟悉程度
		是否可以与信息科有效沟通

3 数据提取

数据提取时应先进行样本数据提取，样本数据验证通过后，再进行全样本数据提取。原则上，样本数据与全样本数据提取规则相同，仅存在数据量的差异。

数据在提取过程中需满足安全性要求，应由各中心内部使用不可逆算法对敏感数据字段脱敏后汇总处理。在数据提取时，建议使用统一的提取格式，采取“csv”格式提取（“csv”格式通用性较好，可兼容多种软件工具，亦无数据量上限，可避免提取过程中数据的缺失）。数据收集过程中，各中心均应在时间允许的情况下进行小样本预提取；也应提供本中心数据提取需求模板，明确数据存储结构。各中心在数据提取时应赋予患者在该院患者唯一识别码（以下简称“患者ID”），并用就诊码（以下简称“就诊ID”）区分同一患者不同就诊记录，通过患者ID与就诊ID链接各表间同一患者的不同字段信息。回顾性数据提取应注意各类数据提取时的要点，具体如下。

3.1 数值型数据提取

数值型数据提取时，应同时提取用以辅助分析的字段，以满足数据转化的要求，如各类医嘱均应同时提取其开立时间、停止时间、执行频次等。用药记录应提取完整的通用名及商品名、剂量及单位、给药途径、给药频次、开始及停止用药时间；一日记录多次的生命体征应同时提取全部记录数值、单位、测量时间。实验室检验检查结果应附量纲及标准值；多中心数据需同时提取各中心的检验检查结果正常值，用于数据转化。若单独标记“标本类型”（如血液、尿液、粪便等），应同时提取“标本类型”中的内容。

3.2 文本型数据提取

文本型数据提取时，应尽可能提取完整的信息。如超声、CT、核磁共振等影像学检查及心电图、肌电图等电生理检查，一般情况下仅提取文字报告，条件允许的情况下应提取图像资料，由专业医师根据研究进行二次识别分析；现病史、病案首页记录、会诊记录、不良反应分析等主观记录材料根据预先设定的关键词从病历记录中提取。

4 数据验证

4.1 样本数据验证

开展样本数据验证时，应优先比较样本数据与数据提取接口文档，查验字段完整性及字段含义准确性。未提取到的数据应核实原因，若是在数据提取时遗漏，应立即补充数据后再进行样本数据验证；若因中心系统建设问题无法提取数据，应考虑采用其他方式补充数据（如按照病历号查找病例并手动补充临床系统中的可见内容），或考虑该部分数据是否为必需提取的内容。多中心的数据提取要求较单中心更严格。单中心数据提取大多只考虑数据完整性，而多中心数据提取要确保提取字段的意义一致，如检验子项“时间”统一为“送检时间”而非“报告时间”。同时，要注释易误解的字段，如总剂量、处方取药、给药停止时间或医嘱停止时间的定义。

4.2 全样本数据验证

数据正式提取后，应首先验证提取出的数据与数据接口文档的一致性及数据完整性。对于过程中出现的数据缺失情况，要如实记录原因。同时也要检查数据结尾是否存在截断情况，如果存在，需考虑数据是否在提取过程中丢失，如果数据丢失需重新提取。由于正式提取的数据量可能与样本数据的量级差距较大，在提取大样本数据时可能出现数据截断，因此并不应默认样本数

据提取完整的中心，在全样本数据提取时的数据也完整。每次提取数据时均应开展数据验证。

5 数据合库

数据合库时，应先标注中心名称（或代码）及各条目数据序号，以便追溯原始数据。按照数据验证结果创建整合表，并从原始表选择插入目标字段、中心名称（或代码）、各条目数据序号。数据合库过程中，应记录整合表字段与原始表字段的对应关系，并检查选择插入的数据量和原始表中的数据量是否一致，避免合库过程中导致的删失。

6 数据核查

开展数据核查时应先将各表通过患者ID及就诊ID进行关联，确定编码映射关系是否对应且唯一，如确定各表间患者数及病例数是否对应。建议使用计数的方式检查每个表的数据记录数、字段数、病例数，检查表间是否存在数据缺失。值得注意的是，提取出的数据中的就诊ID应为脱敏后的患者住院号或就诊ID，与院内人员在本院系统中所见就诊ID不应为同一号码。回顾性研究数据的逻辑核查至少应包括2个方面的内容——完整性、准确性[10-11]。值得注意的是，前瞻性数据核查应注意的内容（如违背方案）本文暂不涉及。

6.1 数据完整性核查

数据完整性核查应检查数据信息的缺失程度，包括变量的缺失和变量值的缺失。首先，应基于源数据适用性评价进行核查，数据应满足相关研究目的所要求收集的最少变量信息，多中心回顾性电子病历数据的源数据各数据表注释及必需字段见表2。其次，应核查单个变量的记录完整度是否满足待研究变量的最小统计效能，在充分考虑混杂因素、缺失数据等因素的基础上满足统计假设的要求。

表2 多中心回顾性电子病历数据的源数据各数据表注释及必需字段

数据表	注释	必需字段
患者表	患者本人的不可变信息	患者ID、性别、出生日期、民族等
病历/就诊表	患者本次就诊信息	患者ID、就诊ID、入院日期、出院日期、就诊类型、入院科室名称、出院科室名称等
用药表	本次就诊的全部用药信息	患者ID、就诊ID、药品名称、药品剂型、药品规格、单次剂量、给药频次、给药途径、用药开始时间、用药结束日期、医嘱类型等
检验检查表	本次就诊与研究相关的检验检查项目的全部信息	患者ID、就诊ID、检验标本、检验检查子项名称、检验检查结果、检验检查单位、检验检查时间等（需另附检验检查子项正常值范围）
诊断表	本次就诊的全部诊断信息	患者ID、就诊ID、诊断名称、诊断时间、是否为主诊断、诊断类型等

6.2 数据准确性核查

数据准确性核查包括一致性、合理性2个方面，具体如下。

6.2.1 一致性核查 一致性核查包括指向一致、定义一致和格式一致[12]。指向一致指相互关联的数据应符合指定约束关系，如通过身高、体质量计算所得体质量指数应与记录值相同。定义一致指不同表内指向同一变量的数据定义及内容应相符，数据与其描述的客观特征应一致，同一变量在不同样本记录中应采用统一的数据定义，如不同中心记录的“药物剂量”“单次剂量”等说法均代表“单次用药剂量”，需统一命名。格式一致指数据字段内容应遵守统一格式，如日期MM/DD/YYYY与DD/MM/YYYY易产生混淆。

6.2.2 合理性核查 合理性核查内容包括时序、阈值范围、逻辑等[12]。合理性核查中，时序核查包括出入院时间、检验检查取样时间、送检与报告时间、医嘱开立与停止时间等。如存在超出研究设定提取时间范围的数据，应与研究人员讨论决定如何处理超出研究时间范围的数据（同时应检查是否存在由于合库导致的时间格式错误）。阈值范围核查，即核查检验检查异常值、用药剂量、用药开始或停止时间等是否合理。值得注意的是，本规程仅涉及核查逻辑不合理的异常值，如应为正值但是记录为负值、应记录为“+/－”但记录为数字或文字、时间在1900年以前或者未来等异常值。逻辑核查范围较宽泛，如结局变量随时间变化趋势是否合理。

7 数据清洗

经核查完整、准确的数据需要采用信息技术方法对其进行汇总、清洗与转化，形成集成数据。在数据清洗或任何涉及改变数据记录的操作时，应记录步骤及具体命令代码以便后续回查。数据清洗包括对原始数据进行重复或冗余数据的去除、缺失值的核查及补齐等。

7.1 重复或冗余数据的去除

对于明确重复或冗余的数据应去除。为提高数据治理的效率，不在研究范围内的信息也应去除。如以“停药”为目的的单独医嘱；与研究无关但同时被提取出的医嘱字段。无法明确为冗余信息的数据，需提前了解各中心和各治疗领域的开嘱习惯。如对领药、退药医嘱的处理，应根据不同研究方案和需求，与研究者确定。如果患者样本量大，用药情况复杂，退药较少，且不考虑合并用药剂量，则退药相关医嘱不应被纳入研究。而研究生物制剂时，常需考虑各大常用药物剂量及频次，所以仅剔除“领药医嘱”，单独讨论“退药医嘱”。

7.2 缺失值的核查及补齐

数据完整性处理应首先检查缺失数据能否补齐。补齐方式应根据数据的缺失程度、缺失原因和变量值的缺失机制设定。如出现关键数据无法补齐的情况，应剔除无法补齐的病例。回顾性真实世界研究中常用的缺失数据处理方法包括完整观测分析、可用观测分析、末次观测值结转法、均数填补等[1]。在填补用药时间时，应考虑目标药品的用法用量。当涉及文字信息的缺失时，可尝试通过其他字段内容填补。如存在用药信息缺失，

可尝试通过医嘱文字识别。也可通过医嘱识别并补齐缺失的手术名称，通过主诉、病史等文字描述补充吸烟、饮酒、药物和食物过敏史等信息。

8 数据转化

数据转化包括诊断相关字段转化、药品相关字段转化、检验检查指标转化、其他变量转化和衍生变量计算等。多中心数据转化的基础是建立诊断标准库、药品名称标准库等通用数据模型，必要时应按照国际临床数据交换标准协会数据标准进行数据的采集、交换、管理、分析和存储[12]。国内尚未有多方认证的标准名称库，因此更应注重标准的积累。诊断、药品、检验检查等需标准化的数据应结合研究目的，提前确定需转化的数据范围，以提升数据的转化效率。

8.1 诊断相关字段转化

诊断字段转化应优先参考ICD编码，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[12]。如诊断字段存在将多个诊断用分隔符链接的情况，可先使用正则表达式等方法识别出单独诊断，之后根据ICD将诊断标准化。

8.2 药品相关字段转化

药品名称标准化时，应将药品的商品名、别名、英文名等统一标准化为（带剂型的）通用名，建议将目标药物匹配解剖学、治疗学及化学分类系统（anatomic therapeutic chemical，ATC）编码[13]，以作用器官或系统、治疗、药理和化学特性将药物分组，生成可比较、可分类的药物编码，方便研究者分析用药模式、进行用药分层或分组等。药品名称标准化时，应注意考虑给药途径，并根据研究目的选择或排除药品。标准化时也应考虑用药频率，注意临时医嘱的剂量记为当日1次，以及临时医嘱与长期医嘱时间重合时的判断和计算。对于首要目的为含药物使用剂量的研究，如需考虑精确的药物使用剂量时，研究成员应讨论后逐一判断。为提高效率，在方案允许的情况下，可将带有“葡萄糖”“氯化钠”等不影响主要成分的注射溶剂，统一标准化为其主要成分的（带浓度、剂型的）通用名。同一研究中，药品剂量单位需统一。

8.3 检验检查指标转化

在转化多中心检验检查指标时，需注意对数据定义、命名及单位的统一，区分不同标本来源的同名检验项目。检验值转化过程中应注意携带量纲，保证数据合并后分析计算时统一单位，并以相同的数据量级展示。检验检查指标标准化基于检验检查正常值范围，每家中心均应单独提供，绝不可默认范围一致。

8.4 其他变量转化

对于具有多个适应证或有除治疗之外用途的药品，应根据研究目的判定是否需要明确处方的用药目的。用药科室可以帮助区分患者，间接判断用药目的。对于需要分析科室的研究，需注意统一各中心科室命名。

8.5 衍生变量计算

涉及用药及检验检查时序判断时，应统一时序计算标准，如患者就诊当日记为第0天。在检验检查结果的判定方面，注意区分包含或真包含（如＜或≤）。此外，检验检查结果异常并不等同于发生不良事件，需使用药品不良反应评价标准进行相关不良事件的判定。

9 数据分析

下文以药品上市后安全性评价为例，描述利用电子病历数据开展多中心回顾性真实世界研究涉及的数据分析方法及要点。

9.1 一般原则

多中心真实世界研究中统计分析的一般原则[1]：（1）所有的统计检验均采用双侧检验，检验水准$\alpha=0.05$。（2）定量指标的描述包括例数、均数、标准差、中位数、第25百分位数、第75百分位数、最小值、最大值等。（3）定性指标的描述包括例数、率、百分比等。

本例中，参考已有文献对开展真实世界研究的基本分析框架要求[1,14-15]，基于多中心电子病历数据的一般情况，将重点从干预性研究转移至回顾性研究，以安全性结局为目的对数据进行分析。生命体征和既往史等回顾性数据缺失程度较大的字段，本例不予强调。利用多中心回顾性电子病历数据开展药品上市后安全性评价可包含的分析内容及要点见表3。

表3 利用多中心回顾性电子病历数据开展药品上市后安全性评价所分析的内容及要点

分析内容	具体内容	分析要点
研究概况	研究人群	对分析人群进行描述，明确人群特征
	数据完整性	根据数据提取需求总结各中心数据提取情况，记录各字段数量
	暴露/纳排情况	（1）队列研究中对观察单位的暴露情况描述。（2）经初步纳排后对样本量的描述
基线情况	人口学特征、生命体征、既往病史、合并诊断等	（1）描述患者特征，衡量各组基线的可比性。（2）对患者疾病严重程度及人群特征进行判断，应注意区分不同中心纳入人群可能存在不同特征，如处方习惯、就诊习惯等，注意结合目标人群患病特征及标准用药模式综合考虑。（3）对患者治疗方案的描述可包括如药物名称、药物剂量、用药频率、时间和疗程的先后顺序等。（4）分析内容包括但不限于人口学信息、基线生命体征、既往病史、生活习惯或个人史、合并诊断、手术史、用药史。（5）可使用倾向性评分法和工具变量法控制基线混杂因素
主要评价指标	研究假设、统计推断、统计模型等	（1）围绕预设的主要评价指标展开，使用单因素分析结合多因素模型，以控制潜在的混杂因素。（2）观察性研究的潜在混杂因素较多，注意进行变量筛选。（3）应考虑可能的交互效应
安全性分析	检验检查结果异常、不良事件（损害）等	（1）不良事件需要描述例次和患者数。（2）分析内容包括但不限于不良事件的诊断或描述、实验室检查结果的变化和影像学检查报告描述的变化等情况。（3）典型药物相关损害应参考目标安全性结局相关解剖系统临床指南或《药品不良反应术语使用指南（征求意见稿）》[16]中不良反应判定评价标准，规定数值标准和有效时限

9.2 数据挖掘

数据挖掘的目标是从大量的真实世界病例数据中发现重要特征的相关关系。自变量和协变量可包括：(1)患者基本信息及生理状态，包括但不限于性别、年龄、烟酒史、家族史、基础疾病、诊断、特殊状态（如妊娠、某些高风险职业等）、基因信息等。(2)对患者生理状态的干预，包括但不限于使用药物、手术及其他治疗手段。上述因素即数据挖掘方法的输入特征（以下统称“特征”）。考虑到真实世界病例数据的局限性，应根据研究目标结局的时效性，选择一次住院或多次就诊可获取的数据，用以观察目标结局。本文仅探讨回顾性电子病历数据的应用，因此此处仅简要叙述适用于回顾性电子病历数据的、挖掘安全性结局相关影响因素的方法，并按照单特征、多特征、多特征序列分类。病例的“全周期”数据往往无法通过单个数据源获取；如需更强的证据来源，应整合涉及患者流转的多中心数据，以完善病例全周期数据。

9.2.1 单特征对结果的影响 常用的方法是通过建立暴露-非暴露对照组（或者A、B药物暴露对照组），对比2组的不良事件发生率，可定性判断某个因素对结果的影响；或可建立低剂量暴露-高剂量暴露对照组，从而定量分析剂量高低对不良反应发生率的影响。这种方法的重点包括建立纳排标准、控制协变量、建立研究基线、选择适宜的方法构建回归方程等。

9.2.2 多特征对结果的影响 多个特征对结果的影响是比较复杂的，原因分析如下：人体病理与干预是一个复杂的因果系统；很多关键事件未被记录在真实世界数据中；无法判断关键事件的实际发生时间。多特征分析常用的方法包括Logistic回归/线性回归（线性、定量）、决策树（线性、定性）、支持向量机模型（非线性、定量）等[17]。通常将数据划分为2～3个数据集，如用于发现因果关系的训练集、用于调整算法参数的验证集（可选）、用于检验算法的测试集等。应重点关注对特征数据的预处理。特征通常分为离散型（如性别、是否妊娠、诊断等）和连续型（如药物摄入剂量）2种数据类型，需要将2种类型归一化，并形成算法的输入特征；对缺失特征应单独处理。

9.2.3 多特征构成的序列对结果的影响 每个特征与结果发生的时间间隔不同，如相对发生时间更早的特征，与不良事件时间上临近的“短期特征”应得到更多的关注。但这并非绝对，由于人体病理和药理的复杂性，较早的“长期特征”亦可能起到更主要的作用。研究者需要从数据中挖掘序列特征对结果的影响。

将时间线上的所有特征和结果按先后顺序构建一个序列，这个序列也是患者病程中的关键控制点。处理序列数据的常用方法是循环神经网络及长短期记忆。这些方法可用于探究序列前后特征与结果的关联，也经常被用于自然语言处理。由于加入了时间属性，此方法对数据要求很高。需要注意的是，真实世界数据中特征记录的时间不一定是特征实际发生的时间，如诊断，患者事实上已经有该生理状态，诊断记录的是该状态被发现的时间。

当使用上述挖掘方法获得的结果不佳时，原因可能有多种，如(1)样本量不够、不均衡或有偏差。(2)特征选择不合理、不全面。(3)特征缺失过多，一些关键数据未被记录在“真实世界数据”中，需要根据实际情况判断原因，从而进一步调整数据或者训练方法。

10 质量控制

数据的质量控制应建立在完善的真实世界研究数据质量管理体系及完善的标准操作流程之上[9]。应特别关注以下几方面——(1)源数据的质量：高质量的研究中心可保障源数据的完整性和准确性，减少数据本身的缺失和偏差，也可提升数据治理的效率。(2)研究方案和数据提取文档的设计：根据数据提取文档采集字段，需确认关键字段已被收集；数据的提取由指定的专业人员按照规定的流程进行提取，非授权人员不应对信息系统有任何操作行为。(3)使用标准化字典：保证数据治理流程记录完整、可追溯。(4)在数据核查、清洗和转化的各步骤都应设有检查文档，避免步骤缺失。

11 数据管理与储存

开展真实世界研究应有独立的服务器用于处理、储存数据。课题原始数据应存储在固定存储设备中；为保障信息安全，应进行异地备份，或至少将原始数据存储在2个固定存储设备中；为满足研究需求，应将多中心数据集中，存储于独立、安全的服务器中，并应分权限管理，具体要求如下。

11.1 数据管理要求

在处理多中心数据信息时，应将文件存储在指定的保密区域中，未经管理员的许可，不得以任何手段将涉密数据信息带出保密区域。在计算机上使用移动存储设备必须有详细的登记和使用记录，包括时间戳与操作内容等。应对研究人员进行Windows域权限管理，严格管理数据的权限访问。存放数据信息系统的服务器仅通过交换机在局域网内部应用，仅允许维护和数据应用的客户端进行访问，其他电脑一律不得访问，以确保系统内数据的安全性。涉及数据储存的软硬件应统一由数据处理部或专人管理，对硬盘进行定期整理，定期保存系统日志。路由器、交换机、数据存储设备和数据库服务器等关键设备应放置在指定地点。应对系统数据实施严格的安全与保密管理，防止系统数据的非法生成、变更、泄露、丢失及破坏。

11.2 数据传输安全

数据需通过信息通道与文件双加密传输，个别情况下使用移动介质传输的数据需对文件进行加密，密钥应

同数据文件分开传送。经远程通信传送的程序或数据（如电子邮件等），须经杀毒服务器扫描确定无毒后方可使用。同时，应对数据传输做好传输记录，通过移动介质完成传输后应清除数据。

11.3 保存形式及存期

不同类型文档应以不同形式保存，数据处理与提取文档、涉及数据的相关文件电子版应永久保存，纸质版应保存5～10年。

12 结论

既有的真实世界研究指南和规范多基于前瞻性研究设计，而针对多中心回顾性研究的指导尚不全面。本研究补充了利用多中心回顾性电子病历数据开展真实世界研究时，现有指南和规范的流程框架在数据提取、治理与分析等关键步骤中所缺失的技术指导，同时针对已有流程框架的实施细节进行描述。研究方案的设计应针对回顾性数据的特点，针对目标结局选择纳入分析的字段；同时应注意该字段是否可在回顾性数据中获取，对于无法直接获取的关键字段，应在研究设计阶段考虑补充获取的办法或使用替代终点。数据提取应建立在医疗人员与数据治理人员的充分沟通和对各中心字段的理解之上，确保表间、中心间字段可链接，并在数据完整的基础上，注重保障多中心间数据字段意义的一致性。对数据的处理应充分利用标准化数据字段库，并结合临床实际情况，任何操作均应记录备案，保障数据治理可溯源。多中心回顾性真实世界数据多源异构性高，应结合研究目的，注重积累不同治疗领域的数据治理要点，构建标准化数据字段库，以充分发挥真实世界数据的研究价值。

参考文献

[1] 王雯，刘梅，何俏，等. 基于常规收集健康医疗数据的上市药品安全性评价研究设计和分析技术专家共识[J]. 中国药物警戒，2022，19(1)：7-12.

[2] 王雯，谭婧，任燕，等. 重新认识真实世界数据研究：更新与展望[J]. 中国循证医学杂志，2020，20(11)：1241-1246.

[3] 国家药品监督管理局药品审评中心. 国家药监局药审中心关于发布《用于产生真实世界证据的真实世界数据指导原则（试行）》的通告：2021年第27号[EB/OL].（2021-04-13）[2022-07-15]. https://www.cde.org.cn/main/news/viewInfoCommon/2a1c437ed54e7b838a7e86f4ac21c539.

[4] 国家药品监督管理局药品审评中心. 关于《药物真实世界研究设计与方案框架指导原则（征求意见稿）》公开征求意见的通知[EB/OL].(2022-07-07)[2022-07-15].https://www.cde.org.cn/main/news/viewInfoCommon/ea778658-adc3d1ae3ffe3f1cc0522e5e.

[5] 聂晓路，彭晓霞. 使用常规收集卫生数据开展观察性研究的报告规范：RECORD规范[J]. 中国循证医学杂志，2017，17(4)：475-487.

[6] 聂晓璐，武泽昊，赵厚宇，等. 使用常规收集医疗卫生数据开展观察性研究的报告规范（药物流行病学版）RECORD-PE规范中文版：上[J]. 药物流行病学杂志，2019，28(3)：190-198，212.

[7] 孙鑫，谭婧，王雯，等. 建立真实世界数据与研究技术规范，促进中国真实世界证据的生产与使用[J]. 中国循证医学杂志，2019，19(7)：755-762.

[8] 吴阶平医学基金会，中国胸部肿瘤研究协作组. 真实世界研究指南2018版[EB/OL].[2022-07-15]. https://guide.medlive.cn/guideline/16342.

[9] 王雯，高培，吴晶，等. 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(7)：763-770.

[10] 于玥琳，卓琳，孟若谷，等. 真实世界数据适用性评价方法的研究进展与前景挑战[J]. 中华流行病学杂志，2022，43(4)：578-585.

[11] 中华中医药学会. 中医药真实世界研究技术规范：数据库构建和数据预处理[EB/OL].（2021-06-30）[2022-07-15].http://www.cacm.org.cn/wp-content/uploads/2021/06/1-中医药真实世界研究技术规范-数据库构建和数据预处理-公示稿.pdf.

[12] 中华中医药学会. 中医真实世界研究技术规范：征求意见稿[EB/OL].（2017-10-16）[2022-07-15]. http://www.cacm.org.cn/2017/10/16/4811/.

[13] World Health Organization. Anatomical Therapeutic Chemical（ATC）classification[EB/OL]. [2022-07-15]. https://www. who. int/tools/atc-ddd-toolkit/atc-classification.

[14] 谷成明，李一，王斌辉. 真实世界数据与证据：引领研究规范 赋能临床实践[M]. 北京：科学技术文献出版社，2022：82-83.

[15] 中华中医药学会. 中医药真实世界研究技术规范：统计分析计划制定[EB/OL].（2021-06-30）[2022-07-15]. http://www.cacm.org.cn/wp-content/uploads/2021/06/2-中医药真实世界研究技术规范-统计分析计划制定-公示稿.pdf.

[16] 国家药品不良反应监测中心. 药品不良反应术语使用指南：征求意见稿[EB/OL]. [2022-07-15]. http://211.166.249.242：8082/xiazaizhongxin/2018-03-16/258.html.

[17] GARG A，MAGO V. Role of machine learning in medical research：a survey[J]. Comput Sci Rev，2021，40：100370.

（收稿日期：2022-08-05 修回日期：2022-09-05）

（编辑：舒安琴）