DOI:10.3969/j.issn.1005-202X.2025.01.018

其他

基于电子数据采集系统关键技术的临床研究数据库的构建

王勇[1],冯前进[1,2]
1.南方医科大学生物医学工程学院,广东 广州 510515;2.广东省医学图像处理重点实验室,广东 广州 510515

【摘要】规范化、高质量的临床研究数据库的构建是临床医师或科研工作者开展临床医学研究的重要环节,是获得高质量科研成果和论文发表的重要保障。本文总结分析了传统临床研究数据库创建的类型及存在问题,重点介绍了基于电子数据采集(EDC)系统进行高质量临床研究数据库构建的关键技术解决方案,包括中央随机与EDC系统的结合、逻辑核查、智能化随访管理、数据字典、痕迹追踪、数据分析与统计化等6大模块的技术创新方案来解决传统软件系统在构建临床研究数据库方面的不足,并对采用EDC系统构建临床研究数据库的应用效果进行了总结与展望,以期为临床医师或科研工作者开展高质量临床研究提供参考。

【关键词】电子数据采集系统;临床研究;数据库
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2025)01-0135-06

Construction of clinical research database using EDC system key technology

WANG Yong[1], FENG Qianjin[1,2]
1. School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China; 2. Guangdong Provincial Key Laboratory of Medical Image Processing, Guangzhou 510515, China

Abstract: The construction of standardized and high-quality clinical research database is an important link for clinicians or researchers to carry out clinical medical research, and also an important guarantee for obtaining high-quality scientific research results and publication. A review on the types and challenges of traditional clinical research database creation is provided, focusing on the key technical solutions of constructing high-quality clinical research database based on electronic data capture (EDC) system, including 6 modules of technological innovation program: the combination of interactive web response system and EDC system, logical check, intelligent follow-up management, data dictionary, trace tracing, statistical analysis of research data, so as to solve the deficiency of traditional software system in constructing clinical research database. In addition, the applications of EDC system in clinical research database construction are summarized and its development prospects are discussed for providing references for clinicians or researchers to carry out high-quality clinical researches.

Keywords: electronic data capture system; clinical research; database

前 言

近年来,随着临床医师和医药科研工作者对临床研究重视程度的增加,国内临床研究项目迅速增多,而临床研究过程中会产生大量医学数据,同时也出现了研究数据质量低下与孤岛化、研究低效重复等问题[1]。临床研究数据库(database)是结构化临床研究信息或数据(一般以电子形式存储在计算机系统中)有组织的集合。具有实现数据共享、数据冗余度减少、数据独立、数据集中控制、数据一致性和可维护性以确保数据安全和可靠等主要特点[2]。

传统构建临床研究数据库的类型主要包括:通过Excel等办公软件构建的数据库(简称Excel数据库),通过开源免费的数据库软件EpiData构建的数据库(简称EpiData数据库)等。Excel数据库具有上手容易、创建简单、数据录入简便等特点,但缺点也十分突出:数据更改不可溯源、数据误删难以恢复、数据库的访问权限密级很低等。EpiData软件支持访问权限控制、逻辑核查、数据导出等比较强大的功能,但EpiData数据库的构建,主要还是采取传统的先以纸质病例报告表(Case Report Form, CRF)收集数据再二次以人工录入的方式,仍然存在容易造成

【收稿日期】2024-10-15
【基金项目】国家自然科学基金(62471214)
【作者简介】王勇,硕士研究生,研究方向:医疗数字化,E-mail: 13926262272@139.com
【通信作者】冯前进,博士,教授,博士生导师,研究方向:医学图像分析,E-mail: Fengqj99@smu.edu.cn

数据缺失、信息滞后、数据采集周期过长等缺点。

为规范临床研究数据库的构建及管理，切实提高科研数据的质量，促进临床研究的发展，应用电子数据采集（Electronic Data Capture, EDC）系统创建临床研究数据库应运而生[3-4]。EDC系统采用国际通用的临床数据交换标准协会（Clinical Data Interchange Standards Consortium, CDISC）[5]的标准来创建数据库，将临床研究从纸质病例报告表（CRF）管理模式转变为以电子病例报告表（Electronic Case Report Form, eCRF）为核心的智能化数据管理模式[6-7]。笔者在应用EDC系统构建临床研究数据库过程中，对EDC系统技术平台提出进一步的改良方案，有效解决了传统构建研究数据库模式的不足。

1 系统的构建

对于药物临床试验，原国家食品药品监督管理总局组织制定的《临床试验的电子数据采集技术指导原则（2016年第114号）》等指导文件[8-9]，详细阐述了EDC系统的基本技术要求，对于研究者发起的临床研究，可参照此规范进行EDC系统数据采集[10-12]。为构建高质量的临床研究数据库，基于EDC系统的技术平台，除了常规eCRF设计外，笔者通过中央随机化方法与EDC系统的结合、逻辑核查、智能化随访管理、数据字典、痕迹追踪、数据分析与统计化6大模块的技术创新方案来解决Excel和EpiData等传统软件系统在构建临床研究数据库方面的不足。

1.1 中央随机

传统随机采用纸质随机信封法，中央随机采用静态随机算法（简单随机、区组随机、分层随机）和动态随机算法，并且随机系统作为EDC子系统融合，及时分配受试者随机号和给药方案，避免出现随机号不一致的问题。

1.1.1 简单随机 简单随机为机器语言最容易学习的抽样随机方法，编码（受试者）被分配至试验组和对照组的概率是相同的，假设事件A发生的概率为$P(A)$，通过配置概率$0\leqslant P(A)\leqslant 1$，即可通过计算机语言完成对每一位受试者的独立分配；假设事件B发生的概率为$P(B)$，A和B是相互独立事件，则$P(AB)=P(A)P(B)$代表事件A和事件B同时发生的概率，通过配置概率$P(A)P(B)$让系统轻松自动完成抽样试验与对照。

1.1.2 区组随机 区组随机表示将样本分配到每个区组内的随机化过程。区组长度可以相同也可以不同。（1）通过配置样本，例如sample size=样本量（用s表示），block=区组块（用b表示），width=宽度（用w表示），以区块样本进行分组随机抽样。如概率：$P(A)=\text{random}\left(\frac{s}{p}\right)$，可配置$P$组内概率（每个区组内试验组和对照组的样本量比例为1:1），提前预判组内分配不均问题。（2）根据项目情况，研究者或相关人员保持盲态情况下，通过系统可采取同一项目设置多个区组长度，以达到研究人员难以预测盲底、减少研究干扰的目的。（3）防止过多碎片组出现，可通过配置固定入组或竞争入组形式，通过计算逻辑实现。

1.1.3 分层随机 如果药物或器械的治疗效应受到一些基线特征影响时，可按特征进行先分层，例如将患者按照年龄分为不同的层次，例如青年、中年和老年），然后层内随机（在每个年龄段内随机分配患者到试验组和对照组中），以保证层内组内均衡。在分层基础上也可再进行区组随机分配，参考（02）区组方式。此随机化方法通过配置样本，例如sample size=140,factor=因素（因素有一定控制，暂设计支持5层），通过系统识别分层因素并计算完成分层随机。

1.1.4 动态随机 即适应性随机，必须通过计算机逻辑语言实现算法，而非固定分配列表。通过样本量和已随机事件对适应性概率计算调整，$P(A)$通过计算算法自动控制调整概率P值，完成概率随机。以最小化算法为例：如样本量s=140，随机概率$P(A)=\text{random}\left(\frac{s}{p}\right)$，随着样本量缩小，$P(A)$会逐渐变大，提前配置区间值，自动计算调整概率值，让概率自动按组回归正常，实际使用中也存在通过阶段调整，达到平衡。

1.2 逻辑核查

临床研究数据变量值（包括单变量和多变量值）的录入可通过逻辑算法在线验证是否正常。通过此模块，EDC系统可实现和优化数据录入时的逻辑控制，达到数据管理与质量控制的要求。数据逻辑核查格式如表1所示。

为了避免缺失值，可将一些重要的变量如性别设置为必填项，如果不填写就无法提交数据；对于年龄、身高、体重、血压、生化等指标，可设置取值范围限制，如将身高变量设置为150~200 cm，不在此范围内者无法录入，并对身高<150 cm或>200 cm的病例进行错误提示，但经研究者判断后仍可录入；多变量关联逻辑，如某患者为16岁且已婚，需进一步核查确认。

1.3 智能化随访管理

对访视时间窗进行智能提醒：（1）距访视时间7天前，首次触发手机短信提醒受试者、研究者/临床协调员；（2）距访视时间3天前，二次触发手机短信提醒受试者、研究者/临床协调员；（3）距访视时间1天前，再次触发手机短信提醒受试者、研究者/临床协调员；（4）超过时间窗提醒。此功能模块通过EDC系统后端程序语言（.php）+服务端语言（.sh）完成智能随访管理：以服务Linux.sh脚本编译，对后端服务Api开放

表1 数据逻辑核查参考格式

Table 1 Data logical check reference formats

字段名	变量名	单位	逻辑判断(核查取值范围)	取值含义(样例)
性别	gender	Male/female	If (gender !="")	Gender cannot be null
年龄	age	years	If (age≤65)	Age 65 years or less
身高	height	cm	If (height≥150 &&height≤200)	Height between 150(inclusive) and 200(inclusive) cm
体重	weight	kg	If (weight≤100)	Weight within 100 kg(inclusive)
收缩压	SBP	mmHg	If (SBP<90 && SBP>139)	SBP value range (Normal value:90-139 mmHg)
舒张压	DBP	mmHg	If (DBP<60 && DBP>89)	DBP value range (Normal value:60-89 mmHg)

接口数据进行自动爬取,形成数据队列依次执行,最终达到对每一个受试者进行访视智能提醒。

1.4 数据字典

针对药物字典和医学字典进行自动检索的数据录入功能将会极大地方便用户的使用。系统自动记录药物和医学字典名称,记录数据版本信息和应用程序编程接口(API)数据处理生成的数据流。通过数据字典子模块,实现在进行受试者CRF数据录入时,可针对WHO Drug药物字典和MedDRA医学术语字典进行自动检索录入。用户在录入关键字、拼音、药物简写或药物英文名称都可直接检索到药物详细数据,并直接选择录入,进一步提高数据采集效能。

1.5 痕迹追踪

EDC系统在构建临床研究数据库过程中,可做到全过程留痕。(1)每一个受试者的数据都有操作日志记录,包括录入者、具体操作、操作时间等,保证每一次录入都能溯源跟踪,任何数据修改都可溯源。(2)对于每一个项目都有独立的详细日志记录,提供更为详细的溯源资料。此功能模块采用数据独立分离技术,对痕迹数据表进行数据视图(view)创建和读取。

1.6 数据分析与统计化

EDC系统对受试者数据和研究进度以自动统计图表形式呈现,包括:入组完成率、随访脱落率、质疑完成率、不良事件(Adverse Event, AE)/严重不良事件(Serious Adverse Event, SAE)统计等,让研究者一目了然掌握临床研究数据质量。研究数据库在锁定后则进入统计程序,为方便统计处理,同时还可以按需导出csv等格式,满足不同用户的统计工作需求。

结合上述6大模块的技术创新方案,笔者主导开发的EDC系统形成了15项主要功能(图1)。基本功能:eCRF设计系统、数据采集统计系统、数据自动核查系统、角色权限分配系统、在线数据导出系统、随机分组系统;特色功能:随访提醒功能、高级搜索功能、数据字典功能、表单动作控制功能、模型设计导入功能;安全陪伴功能:操作痕迹追踪、数据独立分离、eCRF电子签名、双热安全备份。

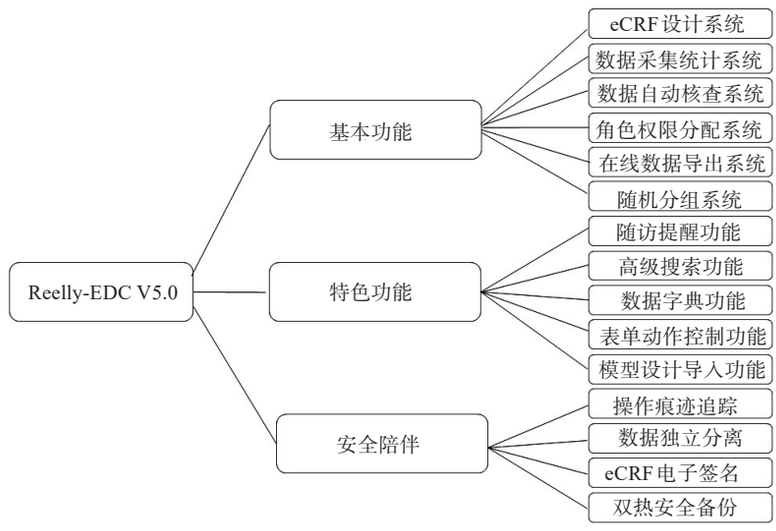


图1 EDC系统主要功能模块

Figure 1 Main function modules of EDC system

2 应用效果

2.1 基于EDC系统构建临床研究数据库的流程

基于EDC系统的临床研究数据库的构建，包括从eCRF建立到数据归档的全过程，主要包括项目建库、数据建模、数据采集、数据质控与核查、数据质疑与清理、数据库锁库与导出等环节流程（图2）。

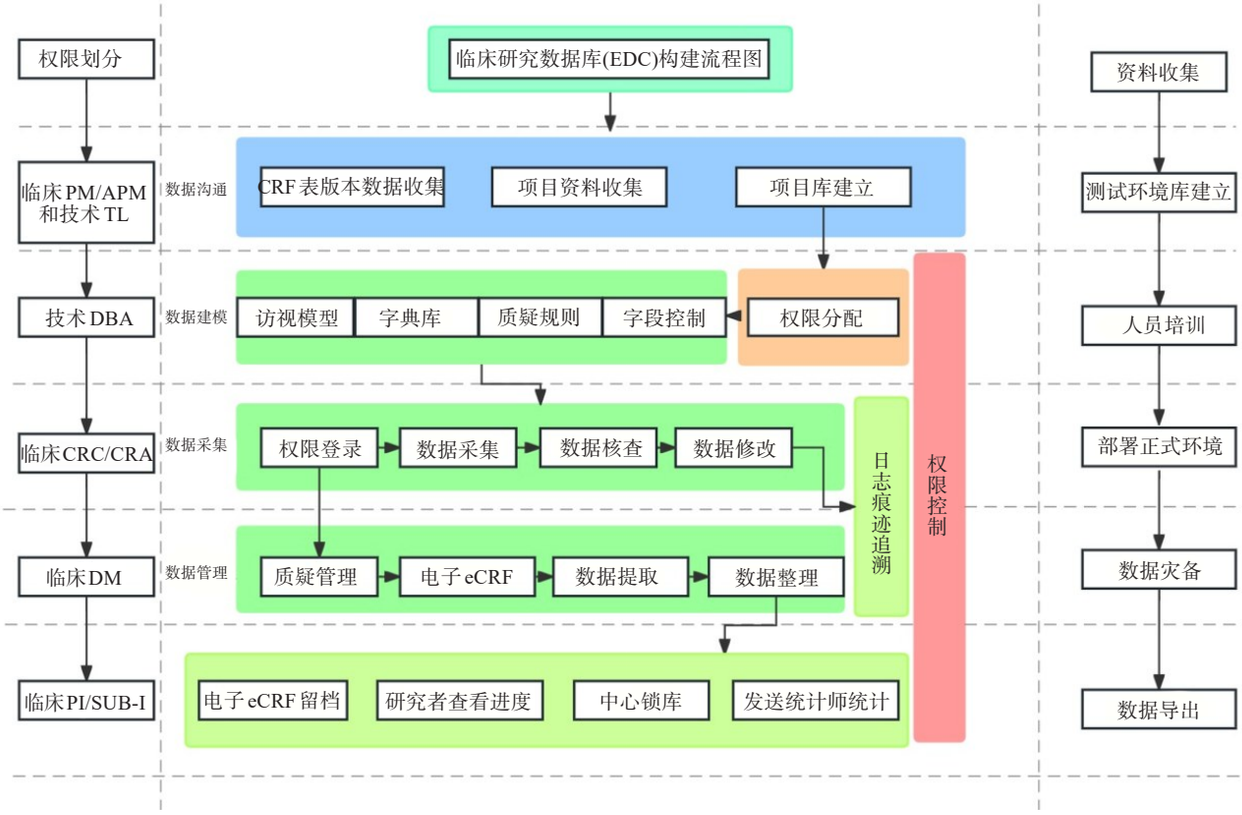


图2 基于EDC系统构建临床研究数据库的流程

Figure 2 Flowchart of clinical research database construction based on EDC system

2.1.1 项目建库 在EDC系统中生成符合具体临床研究方案的eCRF，主要由技术主管（Technical Leader, TL）或数据管理员（Data Manager, DM）负责。CRF（含调查问卷）是记录受试者信息最常用的形式，也是临床研究数据来源的重要载体[13]。建立eCRF测试环境版本后，测试报告反馈经数据管理经理（Data Management Manager, DMM）和项目经理（Project Managerm, PM）审核确认后上线。

2.1.2 数据建模 针对不同临床研究类型项目，根据临床研究方案和标准作业规范（Standard Operating Procedure, SOP），进一步建立访视模型，建立数据字典库，设置数据质疑规则，设置研究团队在EDC系统中的权限分配等工作，然后部署正式环境上线。

2.1.3 数据采集 首次，DM需要依据GCP相关法规撰写项目的数据管理计划（Data Management Plan, DMP），至少包括数据来源、数据字典、研究人员权限、数据提取方式、数据录入方式、EDC操作流程、数据核查计划等。在确定研究方案、CRF定稿和EDC系统上线后，项目管理者需组织对临床协调员进行培训，使临床协调员了解研究目的，掌握数据采集方法和要求以及具体EDC操作系统。申办方/合同研究组织对临床协调员可以制定相应的绩效考核要求，比如数据采集的时限要求、入组工作量、EDC完成率、错误率控制等，以确保数据采集环节的及时性和准确性。

2.1.4 数据质控与核查 数据质控是指申办者派出的临床监查员对EDC采集数据根据研究方案与原始数据进行数据溯源（Source Data Verification, SDV）的过程，以全面督查研究数据的真实性、规范性和完整性。数据核查是指数据管理员根据研究方案要求，对CRF中各指标的数值和相互关系进行核查，进一步完善变量逻辑，以疑问表的形式由临床监查员传递给研究者复核并回答，再对数据库数据进行修订的过程。

2.1.5 数据质疑与清理 数据核查在临床研究不同阶段均进行开展[14]，对于已完成数据录入的EDC数据

库，在正式开展数据统计分析前，由项目数据管理员应对研究数据进行全方面的核查，生成数据质疑，由研究者进行复核和回复。

数据质疑主要包括如下几类质疑工作，(1)录入质疑：输入字段存在错误；(2)缺失质疑：字段存在空白或缺失；(3)异常值质疑：字段值异常(设定了固定范围或固定值)；(4)超窗质疑：访视窗口已超期；(5)数据超范围质疑；(6)人工标注质疑等。项目数据管理员对存在缺失、逻辑矛盾、错误或不能确定的数据发起质疑，对质疑进行处理和关闭，并保留质疑的痕迹。以上数据质疑与清理过程，也是构建高质量研究型数据库的关键[15]。

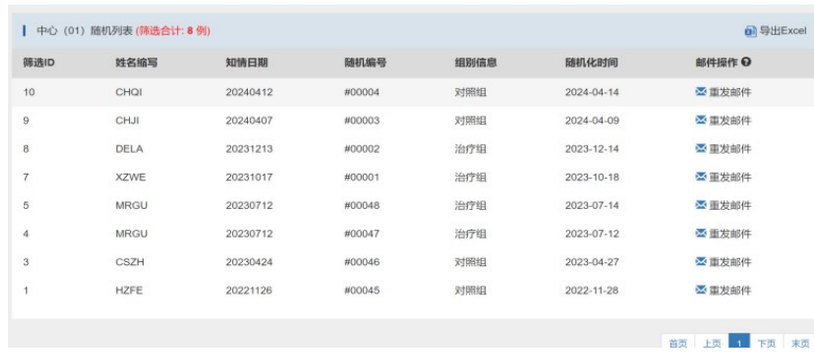
2.1.6 数据库锁库与导出　全部数据审核后，达到统计师分析要求后，主要研究者、数据管理人员和统计师对数据核查报告签字确认，并锁定该项目数据库。数据库锁定后无法对数据进行任何修改，只能将数据内容导出进入统计程序和存档。

2.2 基于EDC系统构建的临床研究数据库的效果分析

以下为基于EDC系统构建的某项糖尿病足临床研究数据库的后台部分效果图，从流程化访视管理、中央随机化、研究数据采集电子化，到对研究进展直观呈现、系统操作留痕记录等，有力保障了高质量临床研究的开展(图3)。

a：流程化随访管理

b：受试者入组列表



c：中央随机化列表

d：系统操作留痕记录

e：研究进展直观呈现

图3　基于EDC系统构建的临床研究数据库后台部分效果图

Figure 3 Some backend interfaces of the clinical research database constructed based on EDC system

研究数据库相对于其他软件系统具有以下强大的优势和特点：(1)规范化。具有标准化的数据采集流程，确保数据的一致性和准确性；(2)数据验证。具备数据验证功能，减少数据输入错误或不完整的数据录入；(3)流程控制。可定制流程，确保数据采集的顺序和逻辑；(4)实时监控。实时监控数据质量，及时发现问题；(5)审计追踪。记录数据的操作变更历史，便于追溯和审查；(6)缩短数据采集周期。减少二次CRF数据录入处理时间，提高工作效率；(7)协作性。支持多用户协作和数据共享；(8)数据安全。使用密码、权限等措施保护患者敏感信息，防止数据泄露；(9)数据集成。可以与其他系统(如电子病历系统)进行接口数据交换，避免重复录入；(10)定制化。可根据具体需求进行定制化功能模块的开发升级。

3 小 结

EDC系统是一个通用平台，可承载各病种数据库，进行各种类型的医学研究和临床试验，包括药物I~IV期临床试验、医疗器械临床试验、研究者发起的临床研究(IIT研究)、患者登记研究、患者随访等。

随着临床试验法规的健全和人工智能技术的普及[16-17]，万物互联发展，未来EDC系统可与终端设备互联，比如：远程访视、语音录入设备、可穿戴手表设备互联等，并可与医院HIS系统打通，实现申办方-医院机构-CRO-SMO多方组织互联互通，可以有效提升研究效率，降低试验失败风险和研究成本。

【参考文献】

[1] 张蒐, 应峻. 临床研究数据管理策略[J]. 复旦学报(医学版), 2017, 44(1): 122-126.
Zhang W, Ying J. Strategies for data management in clinical researches [J]. Fudan University Journal of Medical Sciences, 2017, 44(1): 122-126.

[2] 颜崇超. 医药临床研究中的数据管理[M]. 北京: 科学出版社, 2011.
Yan CC. Data management in clinical research[M]. Beijing: Science Press, 2011.

[3] 王瑾, 汶柯, 王睿, 等. 临床试验电子数据采集系统的国内外现状和发展[J]. 解放军药学学报, 2013, 29(4): 382-386.
Wang J, Wen K, Wang R, et al. Current status and development of the clinical trial electronic data capture system[J]. Pharmaceutical Journal of Chinese People's Liberation Army, 2013, 29(4): 382-386.

[4] 吴泰相, 卞兆祥, 李幼平, 等. 促进我国临床试验数据管理规范化[J]. 中国循证医学杂志, 2018, 18(6): 532-537.
Wu TX, Bian ZX, Li YP, et al. Promoting standardization of clinical trial data management in China[J]. Chinese Journal of Evidence-Based Medicine, 2018, 18(6): 532-537.

[5] CDISC. SDTM[EB/OL]. [2022-01-05]. https://www.cdisc.org/standards/foundational/sdtm.

[6] 齐潇, 肖婉, 陈卉, 等. 标准化数据采集平台构建及在临床试验数据质量控制的应用[J]. 中国临床药理学与治疗学, 2016, 21(12): 1384-1388.
Qi X, Xiao W, Chen H, et al. Construction of standardized data capture system and its application in quality control of clinical trial data [J]. Chinese Journal of Clinical Pharmacology and Therapeutics, 2016, 21(12): 1384-1388.

[7] 房虹, 卢来春, 唐玉, 等. 基于电子源数据到电子数据采集系统的数据直连模式在远程数字化临床研究的应用[J]. 中国临床药理学杂志, 2022, 38(24): 3036-3039.
Fang H, Lu LC, Tang Y, et al. Application of data direct connection mode based on electronic source data to electronic data capture system in remote digital clinical research[J]. The Chinese Journal of Clinical Pharmacology, 2022, 38(24): 3036-3039.

[8] 国家药品监督管理局. 总局关于发布药物临床试验数据管理与统计分析的计划和报告指导原则的通告(2016年第113号)[EB/OL]. (2016-07-29)[2021-12-27]. https://www.nmpa.gov.cn/xxgk/ggtg/ypggtg/ypqtggtg/20160729184001935.html.
National Medical Products Administration. Notice of the general administration on issuing the plan and reporting guidelines for the management and statistical analysis of drug clinical trial data (No. 113 of 2016)[EB/OL]. (2016-07-29)[2021-12-27]. https://www.nmpa.gov.cn/xxgk/ggtg/ypggtg/ypqtggtg/20160729184001935.html.

[9] 国家药品监督管理局. 总局关于发布临床试验的电子数据采集技术指导原则的通告(2016年第114号)[EB/OL]. (2016-07-29). https://www.nmpa.gov.cn/xxgk/ggtg/ypggtg/ypqtggtg/20160729184001958.html.
National Medical Products Administration. Notice of the general administration on issuing guiding principles for electronic data collection technology in clinical trials (No. 114 of 2016)[EB/OL]. (2016-07-29). https://www.nmpa.gov.cn/xxgk/ggtg/ypggtg/ypqtggtg/20160729184001958.html.

[10] 苏毓, 覃开舟, 许健, 等. 研究者发起的临床研究中电子数据采集系统的研究与实践[J]. 中国现代医生, 2022, 60(34): 111-115.
Su Y, Qin KZ, Xu J, et al. Research and practice of electronic data capture system in investigator- initiated trial[J]. China Modern Doctor, 2022, 60(34): 111-115.

[11] 何家双, 肖晓旦. OMOP CDM在临床科研中的应用思考[J]. 中国数字医学, 2016, 11(3): 72-74.
He JS, Xiao XD. Thinking about the Application of OMOP CDM in the Clinical Research[J]. China Digital Medicine, 2016, 11(3): 72-74.

[12] 吕旭东, 田琪, 蔡海领, 等. 临床科研数据库平台关键技术研究与实现[J]. 中国数字医学, 2021, 16(1): 23-29.
Lü XD, Tian Q, Cai HL, et al. Research and implementation of key technologies of the clinical scientific research database platform[J]. China Digital Medicine, 2021, 16(1): 23-29.

[13] 王瑞平, 李斌. 临床研究数据采集策略和要点[J]. 上海医药, 2022, 43(9): 37-42.
Wang RP, Li B. Strategies and critical points of clinical research data collection[J]. Shanghai Medical & Pharmaceutical Journal, 2022, 43 (9): 37-42.

[14] 谭婧, 彭晓霞, 舒啸尘, 等. 患者登记数据库构建技术规范[J]. 中国循证医学杂志, 2019, 19(7): 771-778.
Tan J, Peng XX, Shu XC, et al. Technical guidance for developing patient registry databases[J]. Chinese Journal of Evidence-Based Medicine, 2019, 19(7): 771-778.

[15] 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.
Wang W, Gao P, Wu J, et al. Technical guidance for developing research databases using existing health and medical data[J]. Chinese Journal of Evidence-Based Medicine, 2019, 19(7): 763-770.

[16] 李慧杰, 张晴晴, 刘瑞红, 等. 大数据背景下临床专病数据库建设实践与思考[J]. 中国卫生事业管理, 2020, 37(8): 574-576.
Li HJ, Zhang QQ, Liu RH, et al. Practice and thinking on the construction of specialized disease database under the background of big data[J]. Chinese Health Service Management, 2020, 37(8): 574-576.

[17]《远程智能临床试验专家共识》编写专家组, 上海市药学会药物临床研究专业委员会, 药物信息协会中国数字健康社区. 远程智能临床试验专家共识[J]. 中国新药与临床杂志, 2022, 41(7): 385-392.
Expert Group for Writing the Expert Consensus on Remote Intelligent Clinical Trials, Shanghai Pharmaceutical Association Clinical Research Professional Committee for Drugs, Drug Information Association China Digital Health Community. Expert consensus on decentralized & digitalized clinical trials[J]. Chinese Journal of New Drugs and Clinical Remedies, 2022, 41(7): 385-392.

(编辑：薛泽玲)