真实世界研究的方法学研究成果介绍

海南省真实世界数据研究院北京大学姚晨教授团队

一、 背景介绍

1. 里程碑事件

2016 年 BMJ 杂志上刊登了一起骇人听闻的抹黑中国临床研究的文章“调查发现中国 80%的临床试验数据是造假的”。[1]文中提到：一项对提交中国国家食品药品监督管理总局（CFDA）注册的 1622 个新药数据的调查显示，有 1308 个申请因含有伪造、有缺陷或不充分的临床试验数据而应被撤回。随后，姚晨团队在 BMJ 上发表 opinion，立刻给予强有力的抨击和澄清。[2]中国政府已表示将致力于确保在中国进行的临床试验的质量。2015 年，CFDA 公布了一项数据完整性保护计划，要求试验申请人检查自己的数据。该计划不仅限于研究人员的自我检查。相反，它为研究人员提供了一个机会，可以在没有进一步处罚的情况下撤回向 CFDA 提交的申请，或者提供一个时间窗口来修改数据质量。重新申请时，国家食品药品监督管理总局对数据进行彻底检查。由于监管严格，违法成本高，故意数据造假几乎是不可能的。

这项里程碑的事件中，姚晨团队提出了“提高中国临床试验数据的准确性解决方案”的观点：即应该建立一个独立的平台，用来电子化同步和存储所有试验涉及的源数据。这个平台应主要由医院搭建和管理，同时负责为研究相关人员授权以及决定与 EMR 系统传输数据的范围，确保只有特定试验受试者的特定试验相关数据被同步和存储至这个平台。平台也应包含一个直接收集或者从第三方平台同步随访数据的功能模块，这样临床试验涉及的所有源数据均可以被存储至一个平台。更进一步，平台中的数据还能够依照特定的数据标准比如临床试验通用标准（CDISC）进行重构或格式调整，同时涉及患者隐私的信息也可被移除或者隐藏脱敏。数据将具备共享的条件，可与现有的临床试验电子数据采集系统或其他平台对接，实现数据的直接传输。

2018 年，姚晨团队描述了中国临床试验在数据管理和统计分析的现状与挑战。临床试验产生的数据需要分三个阶段应用：数据收集、管理和分析。基于以上三个阶段，我们了解到随着政府监管部门投入更多资源和努力，未来的临床试验将采用更高的数据管理和技术标准以及最新的统计分析要求，以及在数据采集、管理、分析阶段，做到“准确、规范、完整”。[3]

2. 乐城试点工作构想

2019 年，符祝等人发表在《中国食品药品监管》杂志的《临床真实世界数据用于药品医疗

器械审评审批的探索——海南乐城先行区的实践》文章，描述了在乐城进行试点工作实施路径的初步设想。[4]文中提出了：“探索新的高效、可靠的数据采集模式：建议在医院信息系统之外，建立临床研究源数据集成管理平台，集成新品种开展临床真实世界研究所需要所有 RWD 源数据，探索新的高效、可靠的数据采集模式，尽量利用信息技术自动或半自动对数据清理和研究数据的格式转化与数据标准化，满足注册申报对临床研究数据溯源核查的要求。”

3. 在乐城实践的基础

研究数据真实、准确、可追溯是高质量临床研究的核心要素，也是目前临床研究透明化理念宣传较为薄弱的环节。如何提高我国临床研究数据质量是各方关注的重要问题。课题组在 2019 已经提出了医院临床研究源数据管理平台及源数据管理流程架构。[5]提高研究数据质量的核心环节是促进临床研究源数据的电子化，尤其是需打通临床诊疗数据与临床研究系统的壁垒。课题组成员，董冲亚等人的研究提出了适用于提高我国临床研究数据质量的解决方案，[5]即建立医院临床研究源数据平台，构建临床研究源数据通用管理流程，加强医院临床研究源数据管理。此外，还提出了真实世界研究项目流程。进一步推进临床研究透明化，应深化研究者对透明化理念的全面理解，加强管理者对研究全过程的监督及管理。[6]

晋菲斐在 2019 对北京大学医学部附属的六家三甲医院（三家综合医院和三家专科医院）开展了定性访谈研究，这项研究报告了有关医院各方人员对真实世界数据应用于临床研究的鸿沟、原因和建议的定性访谈结果，为医院更好的利用真实世界数据提供参考。[7]随着电子化数据采集技术的成熟，中国医院的电子病历系统覆盖率已经超过 90%，用于临床研究的数据管理系统也得到广泛应用。但是目前，部分临床研究的数据采集仍然采用人工数据转录誊抄的方式。造成这些问题的原因主要包括缺乏数据互操作性，文本格式的电子病历信息以及管理部门对数据安全性的担忧。更新医院信息系统，提升数据标准和建立独立的临床研究项目源数据管理平台可能是解决当前问题的可行建议。确定原因和针对性的解决方案可能有助于中国临床研究的发展。

在此基础上，课题组成员，晋菲斐等人的研究探索了基于医院信息化的高效可行的临床真实世界数据采集模式探索，[8]并通过一个眼科医疗器械的真实世界研究项目进行验证，对比了应用信息化数据采集模式和传统临床研究数据采集模式的效率与数据准确性。结果发现利用自然语言处理技术对医学文本数据进行处理在效率上比传统人工数据采集有较大优势，虽然自然语言处理也需要依靠人工进行少量文本的标注与校验，但自然语言处理时间比依靠人工录入的时间节省了 90%。

2020 年，姚晨团队发表在《中国食品药品监管》杂志的《利用好真实世界数据生产高质量真实世界证据支持药械监管 》文章，从正确认识、把握重点、迎接挑战三个角度探讨了如何利用好真实世界数据生产高质量真实世界证据支持药械监管。[9]旨在为研究者、申办者、监管者厘清真实世界相关概念上的误区，明确实施中应关注的重点，同时提出现阶段面临的挑战与应对建议。此外，发表在《中国循证医学杂志》的《面向真实世界数据的临床研究数据治理模式选择》文章中，姚晨团队提出了根据数据质量标准（ALCOA+CCEA）对源数据进行治理的管理体系。[10] 针对临床研究数据信息应用现状，同时提出了基于医院电子病历数据开展临床研究的信息安全策略要点。利用医院电子病历数据开展临床研究的信息安全策略应侧重保障电子病历系统信息安全，并在应用过程中充分保障医疗隐私信息安全的情况下，提供可归因、高质量的研究数据，促进数据的合理利用，服务于临床研究。2021 年，姚晨团队针对临床研究数据特征，提出了构建面向真实世界研究的临床研究数据安全等级划分策略。[11]将临床研究数据共划分为五个安全性等级并明确了各等级相应的数据属性和安全策略。临床研究数据规模日益扩大的应用背景下，实现数据安全等级划分及管理是保障数据安全，实现合理利用的必要措施。

二、 成果总结

在海南省博鳌乐城国际医疗旅游先行区的真实世界数据研究项目实践中，海南省真实世界数据研究院北京大学姚晨团队完成了两大真实世界数据研究的方法学研究成果[12-15]。

1. 真实世界数据采集、治理和管理一体化解决工具研究

相较于随机对照试验数据，真实世界数据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。为了解决真实世界数据存在的质量问题，海南省真实世界数据研究院姚晨团队在长期实践中探索出了医院真实世界数据采集、治理与管理的一体化解决方案，并与博鳌乐城临床研究中心和杭州莱迈医疗信息科技有限公司合作开发出了创新型的电子源数据记录（ESR，eSource Record）工具。ESR 的功能设计主要包含：源数据采集、数据提取和治理、与电子数据采集（EDC）和医院 HIS 系统对接三部分。应用自然语言处理技术（NLP）自动填充 eCRF 的技术，ESR 无需耗时且费力的手动数据转录，直接实现数据从 EMR 到 EDC 的电子传输，从而实现高效和可靠的数据采集，有效地降低研究成本。通过创

新性地优化临床研究的源数据采集过程，并遵循 eSource 理念和 ICH/GCP 等原则的设计，ESR 可以满足临床研究数据质量的 ALCOA+CCEA 标准，保证源数据的实时采集、研究数据记录的完整和准确性，同时提高临床医师撰写电子病历的工作效率。ESR 致力于解决全生命周期的数据质量和整合多来源数据，并遵循临床研究的实施标准。通过对接 EMR 和 EDC，ESR 可以灵活应对当前医疗信息水平现状，实施更简单和易于落地推广，具有更高的规范性和和可持续性。ESR 也为目前正在建设的海南省临床真实世界数据研究平台（一期）提供高质量的 EMR 源数据，解决了制约我国进口特许药械项目在乐城国际医疗旅游先行区开展真实世界研究的瓶颈问题。ESR 的登录界面展示和功能介绍，见图 1。

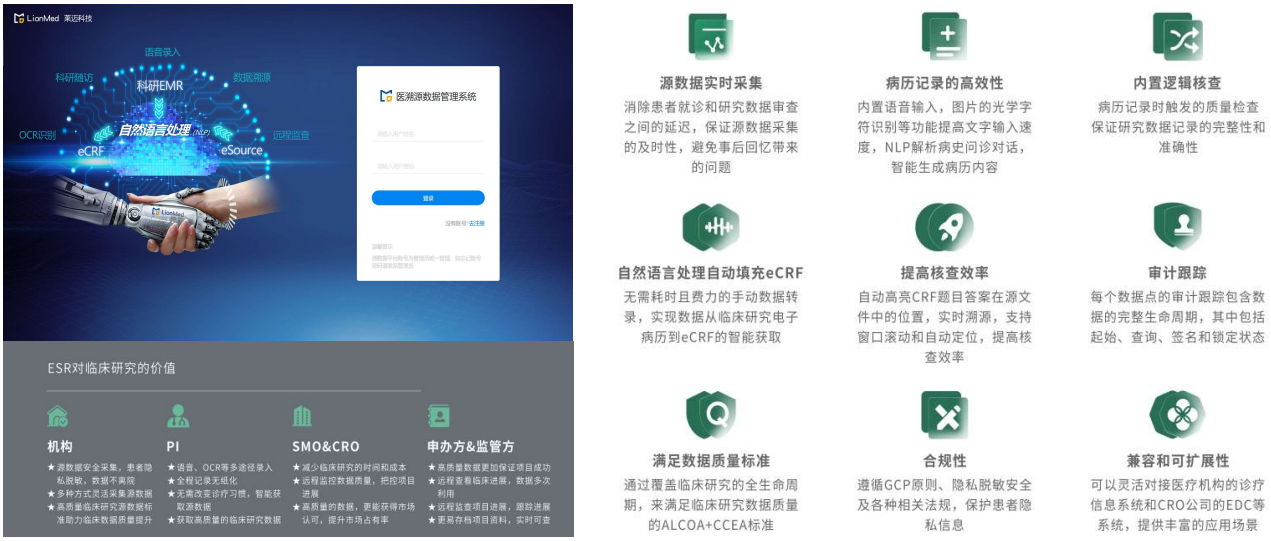
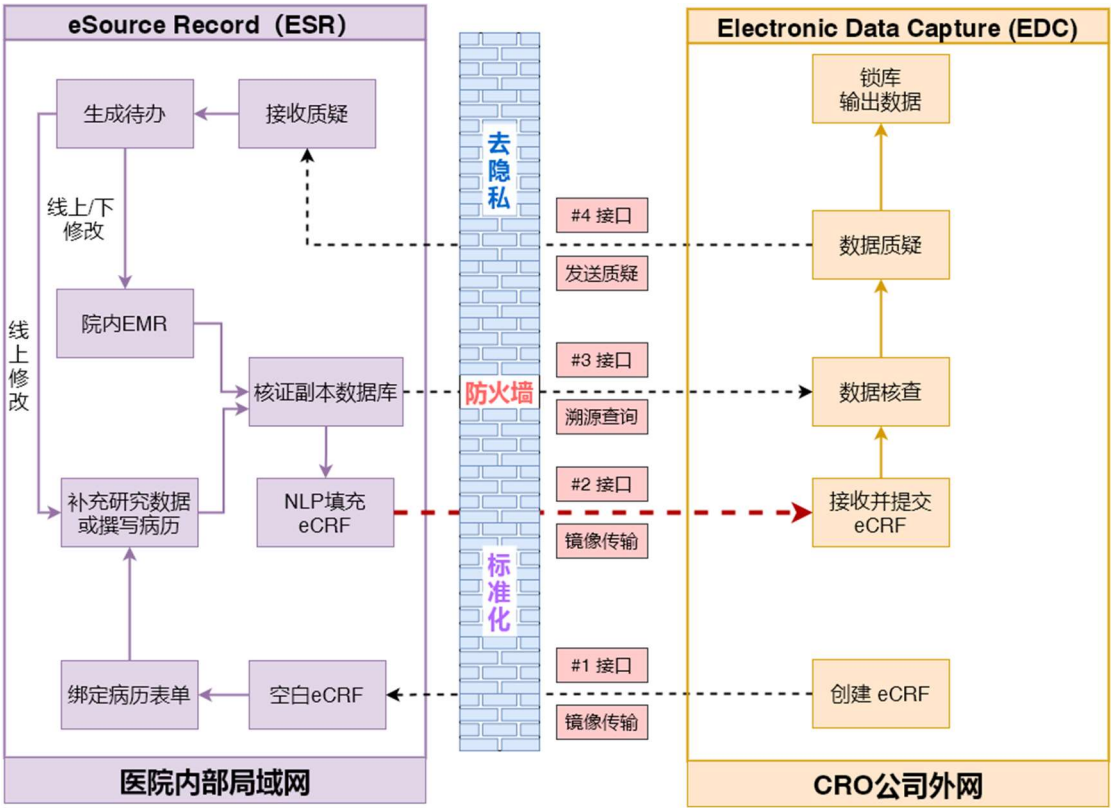


图 1　ESR 的登录界面展示和功能介绍

根据临床试验方案要求，需要将医院已经集成的与电子病例报告表（Electronic Case Report Form，eCRF）中对应的所有电子源数据记录（eSource record，ESR）系统中提取并输出到一个临床试验电子数据采集（Electronic Data Capture，EDC）系统中。临床试验数据自动采集、治理和管理的一体化解决方案可以实现 ESR 与 EDC 的无缝对接，ESR 工具主要涉及 4 个接口的开发：① EDC 创建 eCRF，并通过接口 1 发送给 ESR 工具。② ESR 工具在利用自然语言处理（Natural language processing，NLP）技术自动填充 eCRF 后，将数据通过接口 2 发送到 EDC 系统中，从而在 EDC 系统中进行传统的数据核查和数据质疑流程。③ EDC 系统通过调用 ESR 工具的接口 3，进行数据溯源。④ EDC 系统产生质疑后，通过接口 4 发送到

ESR 工具中。ESR 工具与 EDC 系统之间的对接流程见下图 。



CRO ：合同研究组织（contract research organization）；eCRF ：电子病例报告表（electronic case report form）

图 2 ESR 工具与 EDC 系统的对接示意图

此项研究成果发表在《中国食品药品监管》2021 年 11 月专刊上，见图 3。

图 3 《真实世界数据采集、治理与管理的一体化解决工具研究》

为了评估 ESR 在临床实践中的应用效果，我们在博鳌一龄生命养护中心进行项目试点。选择了一项用于评估美容类医疗器械（GEM HARD 组织修复用材料）的有效性和安全性的真实世界数据研究项目，研究设计为前瞻性、单中心、观察性研究。ESR 系统的 v1 版本于 2021 年 6 月在博鳌一龄生命养护中心进行了部署。经过 2021 年 6 月到 10 月期间使用 ESR 的效果评估，结果发现：以传统临床研究流程作为对比，基于 ESR 的 eSource 方式能提高源数据的采集效率，减少完成数据转录所需的工作量。这项试点的成果证明了 ESR 的可行性和应用价值。整体实践的详细研究成果发表在英国《BMC Medical Informatics and Decision Making》杂志（影响因子：2.796，JCR Q3 区）。[16]研究成果，见图 4。

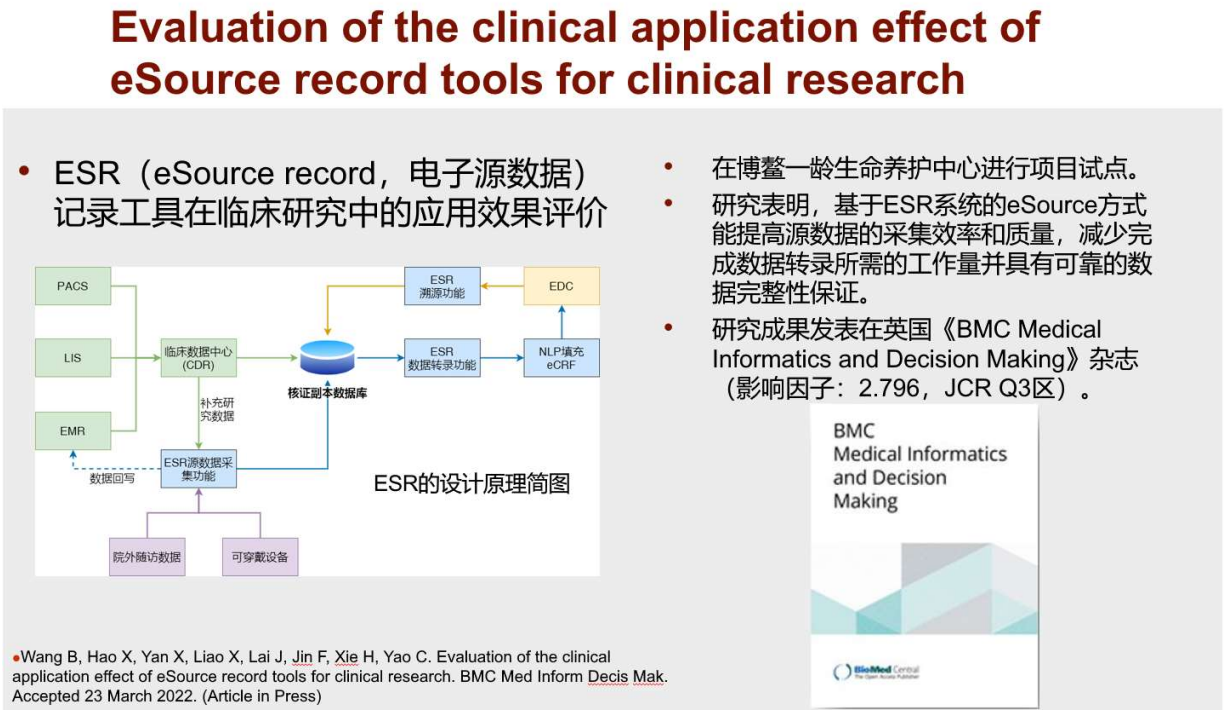


图 4 《Evaluation of the clinical application effect of eSource record tools for clinical research》

2. 从真实世界数据到临床研究数据的标准转化方法

通过定性访谈的方式，对来自 25 个机构（包括 8 家医院、4 家医院系统供应商、3 家大数据公司、6 家制药公司和 4 家监管机构）的 62 名参与者进行了访谈，以调查临床研究中真实世界数据标准化存在的障碍和建议。研究结果表明，现有术语标准缺乏临床适用性，现有研究数据库缺乏普遍性，现有数据标准化过程缺乏透明度，阻碍了这一目标的实现。 通过收集常用术语来扩大术语覆盖范围、减轻术语标准使用负担、利用临床数据模型提高真实世界数据的通用性、提高对源数据的可追溯性以提高透明度，可能是解决当前问题的可行建议。研究成果

以论文形式投稿到《BMJ Open》杂志后，现已收到外审专家的退修意见，在修改补充后送交杂志编辑部终审。[17]研究成果，见图 5。

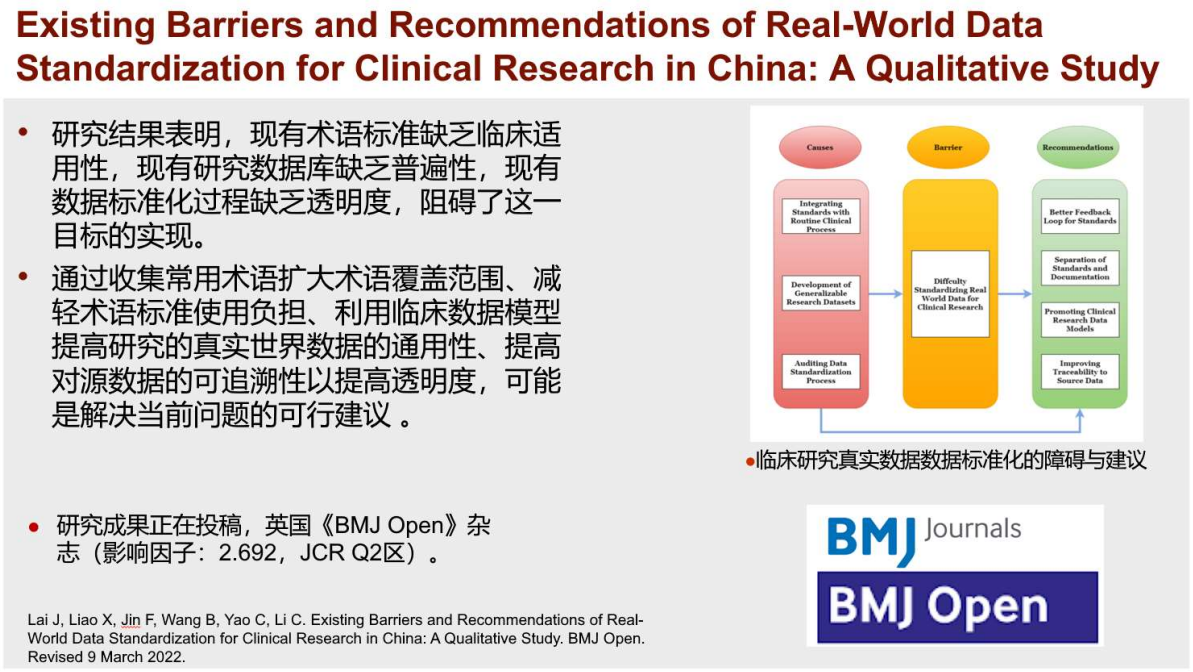


图 5 《Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study》

在分析国内对于 RWD 应用于临床研究的数据标准化需求及挑战后，探索和设计出了一套对数据进行标准化的方法。该数据标准化方法可以用于实现从 EMR 源数据，通过数据标准化的方式自动填充 CDISC 标准的 eCRF，并满足监管部门的数据递交要求。该方法应用了我国常见的数据标准，人工智能领域的 NLP，以及提升数据质量的创新型数据采集模式（ESR 工具）。数据转化过程的核心是根据最简化的数据模型制定文本数据标签指南，提高使用 NLP 算法的效率，优化与临床数据模型的互操作性以及辅助提取研究中所需要的标准术语库。从真实世界数据到临床研究数据的标准化流程如图 6 所示，

该过程包括了以下 5 步：

①由 EDC 发送 eCRF，由 EMR 发送患者临床表单至 ESR 系统；

②对研究数据集建模，生成标签；

③模型训练和实体及实体间关系的提取；

④生成研究专用术语库；

⑤在填充 eCRF 之前对提取实体的规范化规则。

图 5　从真实世界数据到临床研究数据的标准化方法

三、　成果的特色与创新之处

(1) 本研究将首次提出适用于评估真实世界数据整个生命周期所需要应用到的标准和标准化流程，为监管决策提供可供参考的评估工具，将有助于在中国促进区域医院的数据共享和整合，为建设博鳌乐城真实世界研究平台提供指导，最终将以一致的方式提取和治理真实世界数据，完成符合监管规范的真实世界研究项目库的建设，进一步应用到注册研究的审批中。

(2) 本研究首次依据国内 EMR 的特点，开发了一个通用的医疗源数据采集方法的概念框架，适用于临床研究的电子病历文档记录过程，以促进电子病历用于临床研究的书写质量，同时可以提高临床医师的工作效率。整个源数据采集框架内置多种质量控制措施，从而有效地提高病历书写的质量，基于多种数据输入方法的策略，可以减轻临床医师的使用负担，提高工作效率。研究成果可以为解决后期开展研究时遇到的数据质量问题提供新颖的见解，同时为辅助临床医师从 EMR 提取数据和构建专病数据库提供高效的工具支持。

(3) 本研究将带来的改变包括：通过应用 NLP 技术自动填充源数据收集表单，可以解决数据收集困难，耗时耗力的问题；利用溯源“核证副本”数据，可以方便研究者、监管人员人工比对，从而解决无有效手段进行科研诚信监管的难题；提供通用管理流程，实现临床研究数据全程可追溯，从而改善监察工作耗费巨大人力资源成本的现状；“量化” 科研绩效管理的方式将解决科研绩效管理“一刀切”的难题。

四、 参考文献

[1] Woodhead M. 80% of China's clinical trial data are fraudulent, investigation finds. BMJ. 2016;355:i5396.

[2] Yan X , Dong C , Yao C . Protecting the accuracy of clinical trial data in China. BMJ Opinion; 2018.

[3] Yao C. Clinical trial in China: The status and challenge of data management and statistical analysis. J Evid Based Med. 2018;11:3-6.

[4] 符祝, 高国彪, 林尤海. 临床真实世界数据用于药品医疗器械审评审批的探索——海南乐城先行区的实践. 中国食品药品监管. 2019:4-9.

[5] Dong C, Yao C, Gao S, Yan X, Jin F, Zhu S. Strengthening clinical research source data management in hospitals to promote data quality of clinical research in China. Chinese Journal of Evidence-Based Medicine. 2019;19:1255-61.

[6] 董冲亚, 杨莉, 韩鸿宾, 阎小妍, 姚晨. 深化临床研究透明化理念 加强对研究全过程的监督及管理. 中华医学科研管理杂志. 2019;32:146-50.

[7] Jin F, Yao C, Yan X, Dong C, Lai J, Li L, et al. Gap between real-world data and clinical research within hospitals in China: a qualitative study. BMJ Open. 2020;10:e038375.

[8] 晋菲斐, 姚晨, 马军, 陈蔚, 阎小妍, 王斌, 等. 高效可行的临床真实世界数据采集模式探索——海南博鳌乐城国际医疗旅游先行区的实践. 中国食品药品监管. 2020:21-31.

[9] 姚晨. 利用好真实世界数据生产高质量真实世界证据支持药械监管. 中国食品药品监管. 2020:22-7.

[10] 李雪迎, 沙若琪, 姚晨, 晋菲斐, 王熙诚, 阎小妍, 等. 面向真实世界数据的临床研究数据治理模式选择. 中国循证医学杂志. 2020;20:1150-6.

[11] 李雪迎, 王熙诚, 沙若琪, 姚晨, 晋菲斐, 阎小妍, 等. 临床研究数据安全等级划分的初步探索. 中国循证医学杂志. 2021;21:525-31.

[12] 姚晨, 谢红炬, 郝新宝, 谭云, 李玮, 王斌, 等. 真实世界数据采集、治理与管理的一体化解决工具研究. 中国食品药品监管. 2021:62-70.

[13] 赖俊恺, 王斌, 姚晨, 任元凯, 晋菲斐, 王锴. 从真实世界数据到临床研究数据的标准转化研究. 中国食品药品监管. 2021:15-22.

[14] 曹寒, 姚晨, 阎小妍, 于永沛, 尚美霞. 基于博鳌乐城真实世界数据开展特许医疗器械临床研究的设计类型和统计分析方法探索. 中国食品药品监管. 2021:6-14.

[15] 廖茜雯, 晋菲斐, 姚晨. 使用真实世界证据支持全球医疗器械监管决策现状. 中国食品药品监管. 2021:93-103.

[16] Wang B, Hao X, Yan X, Liao X, Lai J, Jin F, Xie H, Yao C. Evaluation of the clinical application effect of eSource record tools for clinical research. BMC Med Inform Decis Mak. Accepted 23 March 2022. (Article in Press)

[17] Lai J, Liao X, Jin F, Wang B, Yao C, Li C. Existing Barriers and Recommendations of Real-World Data Standardization for Clinical Research in China: A Qualitative Study. BMJ Open. Revised 9 March 2022.