真实世界数据采集、治理与管理的一体化解决工具研究

Research on Integrated Solution Tools for Real-World Data Collection, Governance and Management

姚晨*

北京大学第一医院，北京大学临床研究所

海南省真实世界数据研究院

YAO Chen*

Peking University Clinical Research Institute, Peking University First Hospital

Hainan Institute of Real World Data

谢红炬*

海南医学院第二附属医院整形外科

博鳌一龄生命养护中心

XIE Hong-ju*

Department of Plastic Surgery, the Second Affiliated Hospital of Hainan Medical University

Boao Yiling Life Care Center

郝新宝

海南医学院第一附属医院血液科

博鳌乐城泰格临床研究中心

HAO Xin-bao

Department of Hematology, First Affiliated Hospital of Hainan Medical College

Boao Lecheng Tigermed Clinical Research Center

谭云

杭州莱迈医疗信息科技有限公司

TAN Yun

Hangzhou LionMed Medical Information Technology Co., Ltd

李玮

博鳌一龄生命养护中心

LI Wei

Boao Yiling Life Care Center

王斌

北京大学第一医院，北京大学临床研究所

WANG Bin

Peking University Clinical Research Institute, Peking University First Hospital

赖俊恺

北京大学第一医院，北京大学临床研究所

LAI Jun-kai

Peking University Clinical Research Institute, Peking University First Hospital

晋菲斐

北京大学人民医院，国家创伤医学中心

JIN Fei-fei

National Center for Trauma Medicine, Peking University People's Hospital

中图分类号 R95　**文献标志码** A　**文章编号** 1673-5390（2021）11-062-09　**DOI** 10.3969/j.issn.1673-5390.2021.11.008

【摘要】随着网络信息技术的提高，医疗卫生相关数据的电子化已达到较高水平。在传统的研究方式中，通常需要耗费大量的人工和时间成本对相关信息进行整理和提取。而应用自然语言处理技术对诊疗文本进行数据的自动化提取，可明显提高研究效率，降低研究成本。基于真实世界数据的产生和应用场景，以及满足监管部门的法规要求，本研究构建了一个用于真实世界数据采集、治理和管理的一体化解决工具，即电子源数据记录（eSource record，ESR）工具，功能设计主要包含源数据采集、数据提取和治理以及电子数据采集系统对接三部分。为了验证解决方案的可行性，2021 年 6 月，本研究采用 ESR 工具，在海南博鳌乐城国际医疗旅游先行区一家医疗机构的医疗美容科进行了试点研究，对其真实世界研究的源数据进行了采集管理。

【关键词】真实世界数据；源数据管理；数据采集；数据治理；电子源数据记录

[Abstract] With the advancement of information technology, the digitalization of medical and health-related data has reached a high level. Traditional research methods are typically labor intensive and time consuming to sort and extract this information. The use of natural language processing technology to automatically extract data from medical texts has markedly improved research efficiency and reduced research costs. Based on the generation and application scenarios of real-world data, as well as the requirements of regulatory authorities, this study constructed an integrated solution tool, namely the electronic source data record (eSource record, ESR), for real-world data collection, governance, and management. The functional design of ESR mainly includes three parts: source data collection, data extraction and management, and integration with the electronic data collection system. In June 2021, we conducted a pilot study at a cosmetic dermatology department of a medical institution in the Boao Lecheng International Medical Tourism Pilot Zone in Hainan. To verify the feasibility of ESR, we used it to conduct source data collection for a real-world study.

[Key words] real-world data; source data management; data collection; data governance; eSource record

循证医学的兴起使临床医学实践的决策过程从“基于经验”走向“基于证据”。通常情况下，制定卫生决策以及指导临床实践的证据都来于设计严谨的随机对照试验（randomized controlled trial，RCT）。随着人们对疾病认识的不断深入以及现代医学技术的发展，临床实践对多样化证据的需求不断增长，临床疾病诊疗面临的局面也变得越来越复杂。而基于理想情况设计的 RCT 获得的结论往往无法在实际的临床环境中达到满意的效果[1]。因此，基于真实世界数据（real-world data，RWD）的真实世界研究（real-world study，RWS）逐渐被医疗卫生领域所关注。与 RCT 不同，RWS 的研究对象常采用相对较少的排除条件，使纳入的人群有较好的代表性，因此，其研究结果具有更好的外推性[2]。因为 RWS 不是一种特定的研究方法，目前对于 RWS 没有明确的定义。广义上来说，RWS 是通过对 RWD 进行分析研究从而得到真实世界证据（real-world evidence，RWE）的临床研究过程。

1 建设背景

1.1 RWD 的来源

21 世纪是数据的时代，随着深度学习算法的不断发展，基于信息化和行业大数据普及的

真实、高质量的数据是临床研究结果有效、可靠的保证。在长期的实践过程中，监管机构逐步形成了针对临床研究数据的质量标准，可归纳为真实、准确、完整和可靠。

人工智能技术取得了前所未有的成功，这将比以往任何时候都更能解决医疗领域的实际问题。国际药物经济学与结果研究学会（International Society for Pharmacoeconomics and Outcomes Research，ISPOR）和美国食品药品监督管理局（Food and Drug Administration，FDA）对RWD进行了较为明确的定义：RWD除了传统临床试验以外，指从多种来源收集的患者健康状态以及医疗保健相关的数据[3-4]。由于不同的组织机构对RWD的概念存在着一定的分歧，但其核心都在强调数据来源于临床医疗环境，这与医疗实践保持一致。总体而言，RWD包括常规收集的健康医疗数据和基于一定研究目的、进行主动收集的数据[5]。常规收集的医疗健康数据包括医院电子病历数据、电子健康档案数据、院内感染上报数据、药物不良反应上报数据、医保数据等。国外利用RWD进行的研究主要集中在医疗器械[1]，我国则在海南博鳌乐城国际医疗旅游先行区开始进行特许药械方面的探索。

1.2 RWD的采集

随着网络信息技术的提高，医疗卫生相关数据的电子化已达到较高水平。同时，得益于大数据技术和数据安全技术的发展，智能化和信息化的数据采集模式逐渐应用于RWS，且其明显缩短了研究时间，减少了人工重复性的劳动。在信息系统中，除了结构化的数据之外，还包含了大量的文本类型的非结构化数据，其中记录了许多重要的诊疗信息，包括患者的既往病史、症状变化等。这些数据对于临床研究有着非常重要的意义。在传统的研究方式中，通常需要耗费大量的人工和时间成本对这些信息进行整理和提取。而应用自然语言处理（natural language processing，NLP）技术对诊疗文本进行数据的自动化提取，使得研究效率明显提高，且降低了研究成本[6]。

1.3 RWD质量要求

真实、高质量的数据是临床研究结果有效、可靠的保证。在长期的实践过程中，监管机构逐步形成了针对临床研究数据的质量标准，可归纳为真实、准确、完整和可靠。FDA于2018年颁布的数据完整性标准（ALCOA+CCEA原则）[7-8]已被许多监管机构所采用，包括可归因性（attributable）、易读性（legible）、同时性（contemporaneous）、原始性（original）、准确性（accurate）、完整性（complete）、一致性（consistent）、持久性（enduring）和可用性（available）。2020年我国颁布的新版《药物临床试验质量管理规范》[9]（Good Clinical Practice，GCP）中首次增加了源文件、源数据等概念，并要求信息化系统应当能够保障采集数据的溯源，并具有用户权限管理和稽查轨迹功能[6, 9]。

综上所述，基于RWD的产生和应用场景，以及满足监管部门的法规要求，本研究构建了一个用于RWD采集、治理和管理的一体化解决工具，即电子源数据记录（eSource record，ESR）工具，能够满足ALCOA+CCEA标准和GCP的要求，且具有以下优势：①相对独立于院内常规的诊疗系统，且不干扰常规医疗行为。②以服务

临床科研为核心，有效地解决了临床医师的“科研痛点”。③集中管理院内、院外数据，在医疗过程中可以对源数据进行实时、高效的采集。④涵盖数据治理和质量控制等过程（如数据的结构化、标准化等）。⑤具备数据溯源、权限管理及操作留痕等功能，可以实现远程监查。⑥实现个人信息的匿名化处理，有利于个人信息（隐私数据）的保护。

2 建设方法

2.1 技术架构

ESR 工具采用浏览器 / 服务器（B/S）的基本架构，极大地减轻了客户端的部署难度和硬件要求，使得系统更容易在院内进行推广和应用。采用 2 台服务器的设计，1 台服务器用来部署应用系统，同时兼做数据库备份使用；1 台服务器用来作为独立数据库使用，保证了多用户同时访问系统时数据库的硬件输入 / 输出（I/O）能够满足多用户并发的需求。用户可以在个人电脑、平板、手机等任一安装浏览器的端口进行访问。

2.2 功能设计

ESR 工具的功能设计主要包含源数据采集、数据提取和治理，以及与电子数据采集（electronic data capture，EDC）系统对接三部分。针对源数据来源多样性的特点，源数据采集功能可以对接多种数据源。通过对接院内电子病历记录的模板或创建新模板的方式，可以让用户在不更改使用习惯的前提下完成信息的录入。在录入方式上，ESR 工具不仅支持传统的手工录入，同时还添加了语音识别和图片的光学字符识别（optical character recognition，OCR）的功能，极大地提高了用户的工作效率和使用体验。在信息录入过程中，系统内置的数据核查逻辑功能会进行实时扫描，即时对病历中记录的错误信息和缺失的研究数据进行提醒，保障数据质量。在病历记录完成后，根据研究方案预先定义的数据采集要求，系统可以自动识别文本信息，从中提取研究数据到对应的数据元素中，并支持用户对提取结果进行溯源查看。同时系统也会对用户的修改操作进行记录留痕，包含修改人员、修改时间、修改内容等信息，保证数据的可溯源性。

2.3 技术方案

2.3.1 数据采集与清洗

ESR 工具数据主要来源于两部分，一是来源于院内信息系统，二是来源于 ESR 工具的录入。院内来源的数据质量与信息化水平有很大的关系，往往存在着数据缺失、重复、非结构化、非标准化等问题。本研究通过数据接口，将采集到的数据进行清洗处理后储存于 ESR 工具数据库中。与院内来源数据不同，由于 ESR 工具的特点，录入的数据往往已具备较高的结构化和标准化水平，因此不需要进行太多的清洗工作。

2.3.2 数据安全

由于医学信息的特殊性，其数据安全尤为重要，本研究搭建的一体化系统是部署在医院内网，通过网闸与院内信息系统对接，只能通过授权才能访问被允许访问的源数据内容。若需要从外部环境访问 ESR 工具，必须经过医院的虚拟专用网络（virtual private networks，VPN）和堡垒机双重授权后才能进行访问，这样可有效保护数据安全。

另外，不论是院内系统还是 ESR 工具都包含了大量研究对象的隐私数据（如姓名、电话、家庭住址等）。如何在充分保障数据准确、可靠的同时，又保护研究对象隐私信息的前提下，实现数据的可获得性与可利用性，已成为临床研究数据管理的重要目标[10]。针对隐私数据，ESR 工具可通过抑制、泛化、替换等方法进行数据脱敏。首先对直接标识符进行脱敏处理，如将直接标识符假名化、加密、抑制或者屏蔽等；其次，再对间接标识符进行泛化或随机化，最终达到在保证数据可用性

的情况下将隐私信息的匿名化处理。在整个数据采集、清洗、传输及应用的过程中，数据处理方及使用方均遵循相应的法律法规要求，并建立完善的数据隐私安全维护机制。

2.3.3 数据录入

本研究支持多种数据录入方式。除了传统的手动录入外，用户还可以进行语音识别和 OCR 录入。

（1）语音识别录入：《医院信息化建设应用技术指引（2017 年版，试行）》中明确了语音识别技术在医疗领域应用的可行性[11-12]。将语音识别技术运用在电子病历的录入中，可以极大程度地提高医生的工作效率，同时准确地记录原始信息，方便后续的溯源；在发生医疗纠纷时，原始录音可以提供医疗证据。本研究的语音识别技术采用了目前较先进的智能语音交互技术。该技术提供了默认的语音识别模型，同时支持用户自定义上传语料进行模型的优化，既可通过上传音频语料，也可上传文本语料进行语言模型的优化，从而提升了语音识别的准确性。ESR 工具利用语音识别技术实现了 2 个场景的数据采集需求：①传统的电子病历语音录入：需要研究者按照模板，将电子病历的信息通过朗读的方式进行信息录入。在这一步骤中，系统返回的文字信息经过后端的处理，会自动拆分成包括主诉、现病史等字段内容，并填入相应的文本框内。②针对医患对话场景的信息提取：在此场景下，系统采集音频流数据并转化成文字发送给后端，再利用 NLP 技术将文字进行说话人的判别返回给前端，由前端显示医患对话的内容。

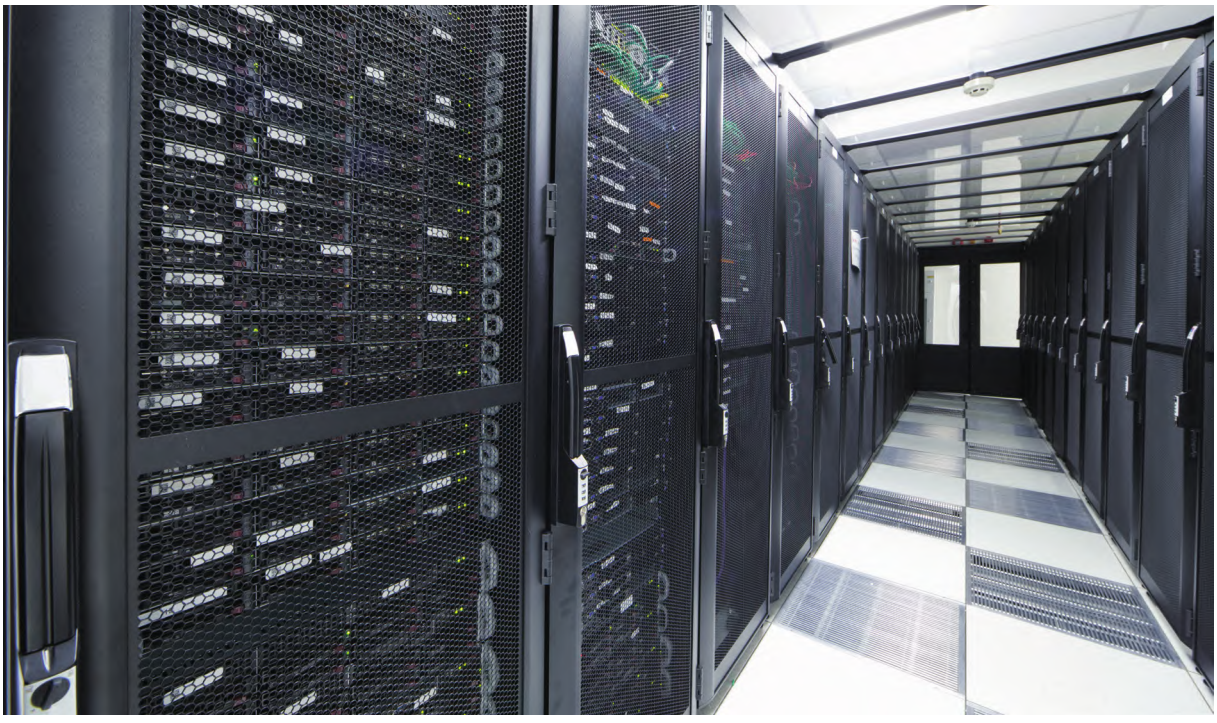
（2）OCR 录入：OCR 是指获取图像文件中的文本及排版信息的处理过程。OCR 具有广泛的应用场景，如车牌识别、身份认证和文档电子化。丰富的应用场景赋予 OCR 技术巨大的商业价值。本研究采用了百度飞浆团队的 PP-OCR 系统。PP-OCR 是一种实用的超轻量 OCR 系统，整个模型大小仅有 3.5M，其不仅在识别表现上较为优异，而且其运行设备的计算能力也很突出。PP-OCR 使用了 MobileNetV3 作为骨干，并采用一系列策略增强模型能力或减少模型大小，其在中央处理器（central processing unit，CPU）上也取得了很好的运行效果。同时，PP-OCR 使用了基于几何中位数的过滤修剪（filter pruning via geometric median，FPGM）算法和参数化裁剪激活（parameterized clipping activation，PACT）量化，在减少模型大小的同时，没有影响到模型的预测表现。PP-OCR 由三部分组成，包括文本检测、检测框校正和文本识别。文检测模块的作用是定位图像中的文本区域，在框选到文本区域后，需要用文本方向分类器对检测框校准，最后对文本进行识别。

2.3.4 数据治理

（1）NLP 技术：NLP 是文本挖掘的研究领域之一，是人工智能和语言学领域的分支学科。主要包括句法分析、信息抽取、信息检索、文本分类、文本生成、对话系统、机器翻译、机器阅读理解等。在本研究中，通过运用 NLP 技术实现非结构化文本的数据自动化提取填充功能，同时辅助语音录入信息的文本转化识别。在实体抽取方面，采用了双向变形编码器（bidirectional encoder representation from transformers，BERT）+ 双向长短期记忆神经网络（bidirectional long short-term memory neural network，BILSTM）+ 条件随机场（conditional random field，CRF）的中文命名实体识别模型[13]。首先，使用 BERT 将中文文本表示成句子的词向量，作为深度学习模型的输入。然后使用双向长短期记忆（long short-term memory，LSTM）模型对输入的词向量进行信息提取。最后使用 CRF 来建模标签

序列，在考虑标签相互关系的同时获得全局最优标签序列。本研究借助于生物医学文本挖掘的双向变形编码器（bidirectional encoder representations from transformers for biomedical text mining，BIO-BERT）的关系抽取模型，将句子中需要判别关系的实体进行替换，然后利用 BERT 模型进行句子向量的表征，借助 softmax 函数的功能输出最大概率，最后判断这 2 个实体是否构成关系。采用和中文命名实体识别相同的框架，利用标注的手段来进行说话人的判别。本方法可以有效地解决由于说话太快，导致语音识别模型无法正常分句的情况。

（2）数据标准化：为提高临床研究数据的质量，需要对临床数据中不标准、低质量的数据进行整合治理，主要包含建立标准化数据集和构建通用数据模型 2 个方面[14]。不同研究者在书写电子病历时使用的词汇往往是多种多样的，包括个人习惯用词、英语缩写等。这些数据在通过 NLP 提取出来后，无法直接被研究使用，需要进一步对数据进行标准化处理。目前，国际上已发表了多种类型的医学术语可供研究使用，如《国际疾病分类》（第 10 版）（international classification of diseases, tenth revision，ICD-10）、《临床医学系统术语》（systemized nomenclature of medicine – clinical terms，SNOMED CT）、《监管活动医学词典》（medical dictionary for regulatory activities，MedDRA）、观测指标标识符逻辑命名与编码系统（logical observation identifiers names and codes，LOINC）等。本研究对诊断、药物、检验、手术等多种类型的数据都进行了标准化处理。在数据库建设上，ESR 系统参考了通用数据模型中的观察医疗结果合作项目的通用数据模型（observational medical outcomes partnership common data model，OMOP CDM）和临床数据交换标准协会研究数据制表模型（clinical data interchange standards



consortium study data tabulation model，CDISC SDTM）标准，使数据能够高效整合、传输和共享。

2.3.5 数据管理

（1）数据溯源核查：研究者在将诊疗数据录入到临床研究电子病历前，ESR 工具已将电子病历与病例报告表做了映射关联，因此研究者在完成数据的录入后，对应的病例报告表能够自动获取数据。此方式极大地减轻了临床研究协调员（clinical research coordinator，CRC）的数据录入工作，使其将主要的精力聚焦于数据核查工作。ESR 工具为 CRC 提供了一种方便快捷且交互良好的数据校验方式，即针对病例报告表中的每一个数据提取结果。如图 1 所示，将鼠标悬浮在“面部美容治疗类型”的结果上面，在左侧原始病历中的信息可以自动高亮。

本研究中用于数据抽取的模型训练的主要算法是 CRF，其中包括词性标注、分词、命名实体识别等领域。首先定义变量数据的抽取指南，将训练文本中的句子进行原子切分，对字（词）进行实体标注，确定特征函数，训练 CRF 模型参数，其中文分词的结构图如图 2 所示。然后将提取的数据进行处理及修正，不断地利用交叉验证原理调整模型参数，以得出最优解。之后将训练好的模型应用于测试集上进行预测，根据模型在测试集上的表现来选择最佳模型。最后运用建立好的模型，选择合适 Pipeline 控件，通过参数调优及特征选择，从数据集中提取变量，形成结果。

（2）用户权限设置：本研究中为保证用户使用数据的严谨性，定义了多维多层级的用户权限设置。在系统层面，可进行用户的增删改查（CRUD），但需要在绑定用户角色后才能进行系统应用。用户角色可自定义，ESR 工具提供了从 1 级到 2 级菜单，再到子功能的多级操作选择。在业务层面，为了方便不同类型的用户应用，在项目维度内制定了主要研究者、执行研究者、临床研究协调员 3 种角色，便于用户在临床研究项目进行时进行任务分配及功能划分。

2.3.6 ESR 工具与 EDC 系统的对接

为了将医院电子病历（EMR）系统数据提取到 EDC 系统，省

图 1　数据溯源示例

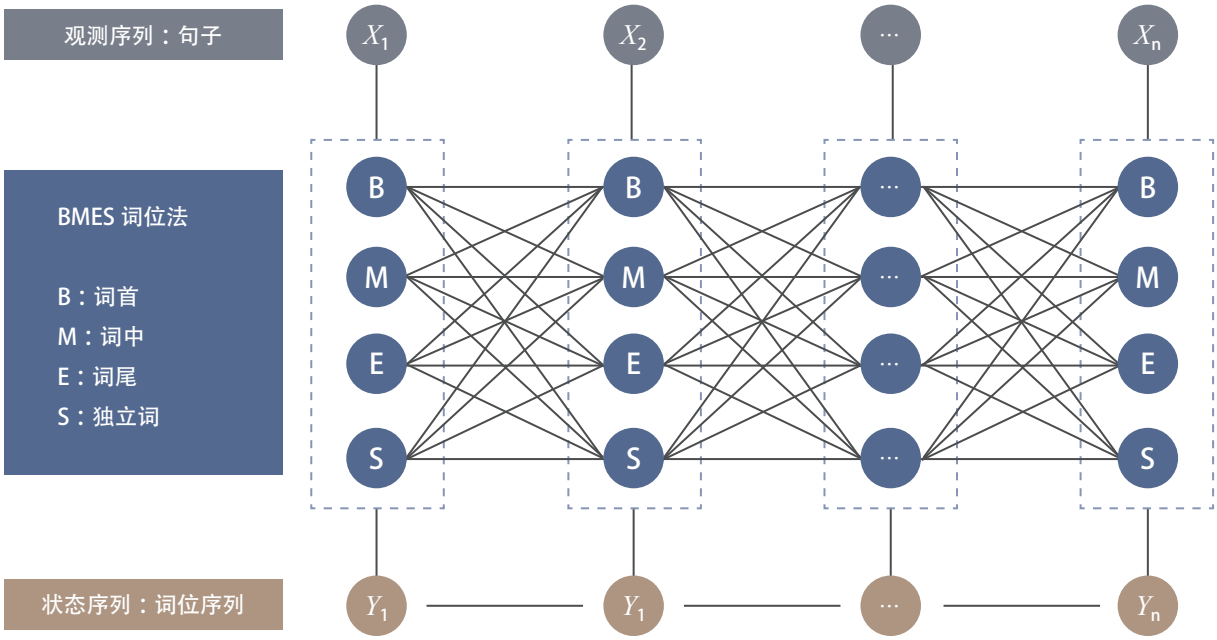
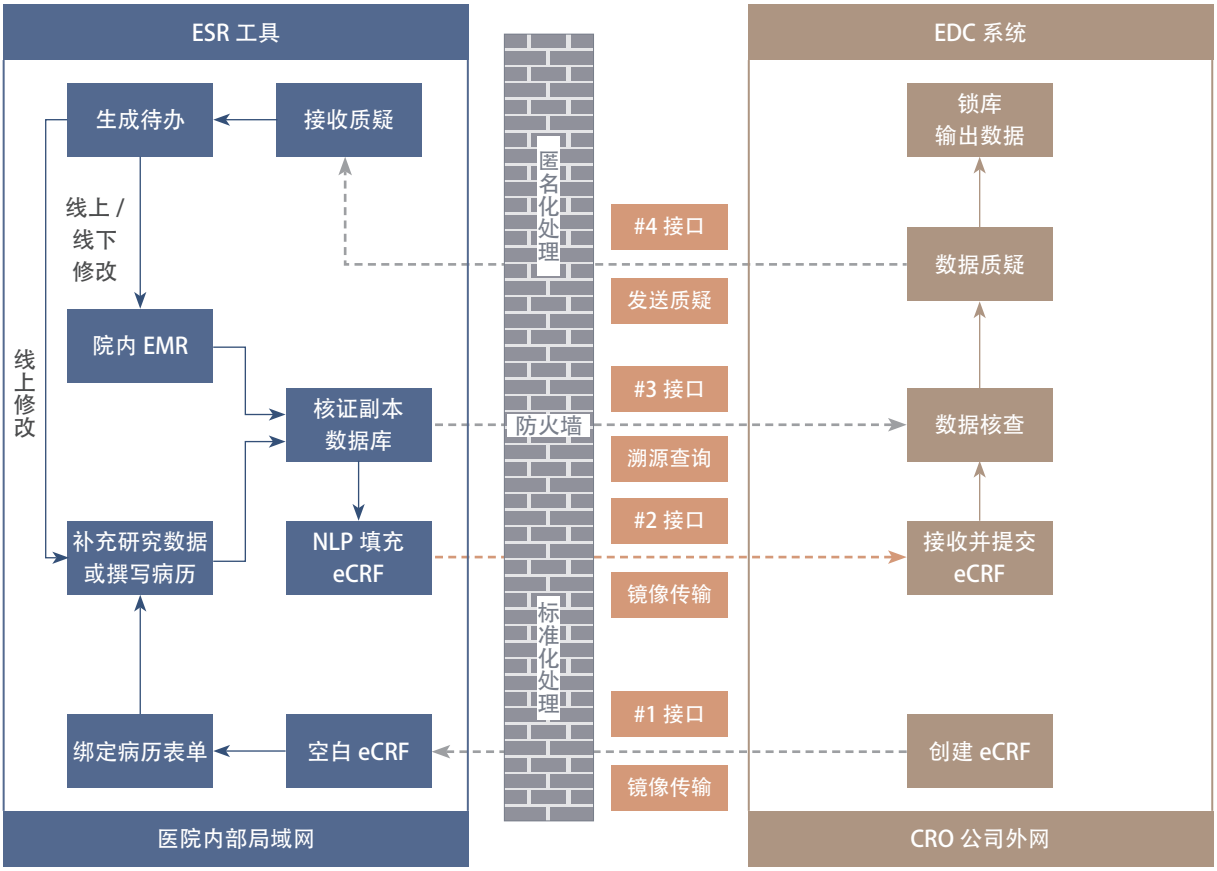


图 2　CRF 模型中文分词结构图

去中间人工手动录入病例报告表的过程，本研究提供了可扩展的系统功能，实现了其与 EDC 系统的对接，从而实现临床研究全流程、智能化、一体的解决方案。2 个系统的对接主要涉及 4 个接口的开发：① EDC 创建电子病例报告表，并通过接口 1 发送给 ESR 工具。② ESR 工具在利用 NLP 技术自动填充电子病例报告表后，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。两者之间的对接流程见图 3。

3 案例应用

2021 年 6 月，本研究在海南博鳌乐城国际医疗旅游先行区选取了一家医疗机构的医疗美容科，使用了 RWS 项目进行了源数据采集工具的试用，来评估该解决方案的可行性。本项目的研究设计为前瞻性、单中心、观察性研究。在项目启动前，首先将源数据采集工具部署到院内，同时与医院的信息科进行了对接。然后对参与本项目的成员进行了系统的使用培训。在项目准备阶段，根据病例报告表定义的数据采集点，配置了每个数据采集点下面临床医生需要完成的病历表单和完善了病历撰写的模板。临床医师综合运用源数据采集系统的多种功能，如语音输入、NLP 解析和自动生成病历等，完成源数据采集。对于研究所需的院外数据，借助微信公众号推送随访问卷的形式进行采集。在完成 ESR 工具与 EDC 系统的接口开发和系统对接后，邀请了从事临床研究的不同成员角色，对应用 ESR 工具的研究全流程进行了正式测试，广泛收集使用者的反馈和经验建议，用于后续系统的迭代升级等。临床医生用 ESR 工具撰写 100 多份电子病历后，评价发现其在病例撰写速度、内容完整性和规范性方面与之前相比有很大提高，同时也满足了该项目的真实世界研究源数据的溯源要求和研究数



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

图 3　ESR 工具与 EDC 系统的对接示意图

据的自动获取，极大地提高了临床研究的工作效率和数据质量。

4 结语

目前，RWD 已成为我国医疗卫生行业的关注焦点，得到了有关部门的重视和推动。2020 年，国家药品监督管理局先后发布了《真实世界证据支持药物研发与审评的指导原则（试行）》[15] 和《真实世界数据用于医疗器械临床评价技术指导原则（试行）》[16]，并于同年启动了海南博鳌乐城国际医疗旅游先行区临床 RWD 应用试点工作。而关于 RWD 的质量管理、安全和隐私保障、统计分析方法，以及由此获得真实世界证据的循证医学证据等级等问题，仍需要多方专家学者进一步讨论完善。本研究通过信息化手段和人工智能的相关技术，搭建了一个用于 RWD 采集、治理与管理的一体化工具，为 RWD 的临床应用提供了一种更高效、更安全、更智能的解决方案。

（编辑：赵文锐）

作者简介

姚晨，卫生统计学硕士，教授，临床研究方法学博士生导师，北京大学第一医院医学统计室主任，兼任北京大学临床研究所副所长和海南省真实世界数据研究院副院长。专业方向：临床研究统计设计与分析

谢红炬，教授，医学硕士，主任医师，硕士生导师，海南医学院第二附属医院整形外科主任。专业方向：美容外科、微整形、抗衰老面部年轻化

参考文献

请扫描二维码