临床医疗大数据治理和应用

王 强 易应萍

（神州数码医疗科技股份有限公司 苏州 215000） （南昌大学第二附属医院 南昌 330006）

〔摘要〕 OMOP通用数据模型是由OHDSI提出的国际先进的医学信息科研数据模型。本文介绍对医院大量临床医疗数据抽取、清洗，在引入OMOP模型的基础上生成医疗科研数据模型，以建立包括精确队列筛选和队列分析、比较等功能的Vinci医疗数据科研分析平台为例，通过平台应用案例，讨论应用意义和存在的问题。

〔关键词〕 临床数据；数据治理；OMOP数据模型

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2018.08.001

Governance and Application of Big Data in Clinical Healthcare *WANG Qiang, Digital China Health Technologies Co., Ltd, Suzhou 215000, China; YI Ying-ping, The Second Hospital Affiliated to Nanchang University, Nanchang 330006, China*

〔**Abstract**〕 The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is a internationally advanced data model in medical information research which is proposed by The Observational Health Data Sciences and Informatics (OHDSI). The paper introduces how to conduct data extract and cleansing on the huge amounts of data in clinical healthcare of the hospital, the medical research data model is created on the basis of introduction of OMOP model. Thereby the Vinci healthcare data scientific research and analytic platform is built, which featuries several functions such as accurate queue filtering, queue analysis and comparison, etc. Through examples of platform application it discusses their significance and issues existed.

〔**Keywords**〕 Clinical data; Data governance; OMOP data modle

1 引言

近年来中国医疗信息化的发展已经步入成熟阶段，目前几乎所有的医院都建立了适合自身需求的信息系统。在各种类别的医院信息系统（Hospital Information System，HIS）中，海量的临床医疗数据被存储在硬盘的一个小的单元中，成为历史医疗数据。随着计算机处理性能的极速发展和单位存储硬件价格的下跌，治理、应用历史临床医疗数据的成本也在快速降低，这为利用数据进行科学研究分析奠定了硬件基础。由于各家医院使用的信息化系统并非一致，在同一家医院内不同的业务模块使用的信息化软件也不相同，为综合研究各医院信息系统，检验信息系统（Laboratory Information System，LIS），医学影像存储与传输系统（Pictures Archiving and Communication System，PACS）等存储的临床数据，通用数据模型（Common Data Model，CDM）[1]的使用就变得非常重要。基于通用数据模型，不同系统、平台、软件中的临床数据被从相应的信息系统中抽取出来，以相同的数据结构存储在数据模型中，研究者可以通过统一的调用方式对不同医院、

〔收稿日期〕 2018-07-17

〔作者简介〕 王强，硕士，工程师，发表论文1篇；易应萍，研究员。

科室的临床数据进行调取、统计、分析等操作。

临床科研是提高医院核心竞争力的基石，是推动医院学科建设和发展的关键，是医院培养和造就人才的必要手段，但目前医院信息系统中临床数据存储分散、缺乏标准化、数据不完整、存在非结构化数据等问题造成临床研究者在科研实践中获取数据时存在种种困难[2]。科研分析平台可以帮助临床研究者快速准确地建立符合自身科研需求的患者队列，进行队列分析或比较，从而更加高效地进行临床科研工作。

2 科研数据模型建立

2.1 数据清洗

2.1.1 概述 医疗大数据是指临床医疗中所产生的海量数据，包括电子病历、医嘱、检查、检验等数据，具有超量（Volume）、多种类（Variety）、高速产生（Velocity）以及真实性（Veracity）4个特征，医疗大数据的规模庞大到完全超出传统数据管理工具的处理能力，无法用传统手段在合理时间内进行撷取、管理、分析。同时由于各个医院信息化建设水平不统一，以及医院内部信息化执行标准不同和信息系统的更新换代，其中的数据存在缺失、重复、不准确、非结构化文本、标准性差、格式混乱等问题。数据清洗的目的就是利用先进的技术手段对医疗大数据中存在的各种问题进行处理，达到补全数据、剔除重复数据、校验数据、从非结构化文本中提取关键数据、数据标准化和格式统一等目的，最大限度利用医院已有临床数据，为科研分析提供坚实的基础。

2.1.2 数据缺失 是在清洗医院数据中最常遇到的一种数据问题，其中又包括两种情况，即可获取数据缺失和不可获取数据缺失。可获取数据是指临床数据中一些客观数据，可以从数据库其他表的字段中通过表关联、计算、推导等技术手段重新获取。不可获取数据主要是指临床数据中一些主观输入的数据，无法通过表关联等技术手段从数据库其他表的字段中获取。可获取数据如年龄、性别数据可以从身份证号码中提取；身份证号或是医保号码可以通过表关联以往的就诊记录中的医保或身份证号，对缺失的数据进行补全；用药天数缺失可以通过药品包装规格、用药频率和完整用药记录进行推导。

2.1.3 数据重复 临床大数据清洗中遇到的又一常见问题，如同一患者拥有多个就诊号码，为在后期数据分析中更准确地对患者信息进行分析，必须将拥有多个就诊号码的同一患者进行统一。利用医保或身份证号对多个不同就诊号码的患者进行统一处理是一种既简洁快速又准确可靠的处理方法。

2.1.4 数据混乱 尽管HIS通常已经对诊断等重要数据进行结构化处理，但是依旧存在很多数据混乱方面的问题，如医生习惯将多个诊断写在同一个单元格中，用空格、逗号或是分号进行区分；或是由于临床医生对国际疾病分类（International Classification of Disease，ICD）编码体系不熟悉，对诊断的描述未使用字典表中的诊断名称，在缺乏有经验的编码员时，难以进行标准化编码；此外国内目前并行存在多套基于国际ICD10标准编码的扩展码，不同医院信息化系统采用的扩展码体系不同。针对数据混乱的问题，需要根据具体情况对同一单元格内的多个诊断利用分隔符进行拆分，或使用字典表进行匹配拆分。模糊匹配评分、搜索评分等方法经常被用来对未使用标准字典诊断名称的诊断数据进行标准化处理，根据标准ICD10编码来映射诊断编码。

2.1.5 非结构化数据 医院的电子病历中存在大量的自由文本数据，如入院记录、手术小结、影像报告等。对于这些非结构化数据，需要采用医学自然语言处理来对关键数据进行结构化处理。

2.2 数据抽取

由于医院临床数据分散存储在多个不同的信息系统中，数据的抽取过程不可避免地与医院多个临床系统产生交集。通过中间表视图，医院临床业务系统中的数据以符合观察性医疗结局合作项目（Observational Medical Outcomes Partnership，OMOP）需求的格式呈现出来。中间表是指连接OMOP通用数据模型和医院临床业务系统数据内容的数据库表。数据库中的视图是一个虚拟表，可以提供和真实表相同的数据内容和字段。中间表视图能够在几乎不增加数据库负担的情况下实时、准确地将科研所需的临床数据从不同的临床系统中以满足OMOP

数据需求的方式查询显示出来。Kettle是数据抽取过程中常用的提取、转换、加载（Extract、Transform and Load，ETL）工具，通过可视化的方式，提供便捷高效的数据提取方式。通过在Kettle中配置输入输出数据库的接口和数据表，设置互相映射的字段，中间表视图中的临床数据被完整地抽取到OMOP数据表中。数据抽取过程使用具有自主知识产权的数据脱敏技术，以保证用于科研的数据经过绝对脱敏且不可追溯原患者，从而保证科研的客观性和患者隐私的保密性。

2.3 通用数据模型建立

医学领域的标准化问题一直是国内外共存的难题，为有效提高后续数据的分析质量，将临床医疗数据转化成研究用的数据模型是当前普遍接受的方法。通过采用由观察性健康数据科学和信息联盟（The Observational Health Data Sciences and Informatics，OHDSI）提出的国际先进的OMOP通用数据模型，很好地解决临床医疗数据的标准化存储问题[3]。该模型允许对不同的临床医疗数据库进行系统分析[4]。将这些数据库中包含的数据转换为通用格式（数据模型）以及通用表示（术语、词汇、编码方案），然后使用标准分析程序库进行系统分析，根据通用格式编写[5]。采用OMOP通用数据模型，医院临床系统中的临床医疗数据通过清洗、抽取等步骤进入标准的存储模型，为后期高效利用临床数据进行科研分析提供基础[6]。

3 科研平台建立

3.1 概述

以神州数码医疗科技股份有限公司的Vinci科研数据分析平台为例，其为建立在OMOP通用数据模型基础上，集队列筛选、单队列分析、多队列比较、运营统计、搜索引擎、样本预测、科研管理等功能为一体的综合医学临床数据科研分析平台，遵循安全性、准确性、可靠性、可扩展性、开放性、实时性、易用性、易维护性等设计原则。基于浏览器/服务器（Browser/Server，B/S）架构，尽可能减少客户端的运行成本，用户只需要使用安装有网络浏览器的计算机、平板或手机就可实现对平台的访问、操作。同时平台的部署完全处于医院内网，既保证安全性，也保证医院数据的使用范围。平台采用前沿的前端网页技术，将原本枯燥乏味的队列筛选操作转换成通过拖拽操作即可轻松实现的复杂关系下的队列筛选，使在紧张工作环境下的医务人员可以享受到轻松的科研工作过程。多种可视化显示技术的使用可以使分析结果进行多元化展现，分析数据更加直观形象。

3.2 主要功能

平台集成T检验、卡方分析、单向方差分析等临床研究常用统计检验方法，同时平台支持样本信息表导出功能，对于有更高分析需求的科研人员，在申请样本信息导出权限后，可将通过队列筛选和添加分析变量后的样本信息以数据表格的形式导出，用于在更专业的分析统计工具中进行分析。搜索引擎是Vinci平台的一大亮点，通过自有医学本体库建立的分词标准，为OMOP通用数据模型中的所有临床数据建立索引，通过搜索引擎可快速、准确地搜索到数据库中所有与关键字相关结果并按评分排序。搜索结果以患者为主索引，列出包括搜索关键字在内的该患者所有的临床信息，以时间轴的方式呈现给用户。平台提供的运营统计功能对整个数据库中的数据进行描述性统计，以多种图表的方式呈现，使用户可以快速了解整个数据库中的数据总体情况；样本预测功能不仅提供样本的快速预测，也提供基于预测模型的结果预测；科研管理功能对科研项目的申请、共同科研、数据使用、数据导出等功能进行权限控制、审批等操作。

4 应用案例

4.1 研究目标

收集并比较因子宫良性疾病需行子宫切除术的患者的术中和术后情况，对两种子宫切除术的临床应用加以探讨，了解不同方式子宫切除术的临床价值，为选择安全性高、创伤小又符合病情的手术方式提供建议。

4.2 研究人群

（1）年龄。18 周岁及以上。（2）诊断记录。诊断为子宫良性疾病，且无心、肺、肝、肾等疾病。（3）手术记录。行子宫切除术，包含腹腔镜全子宫切除以及开腹式全子宫切除。

4.3 研究流程（图1）

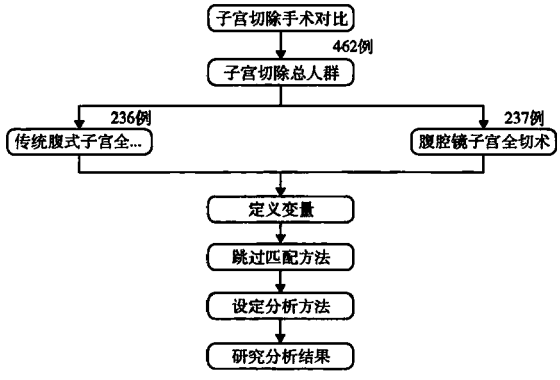

图1 Vinci 科研分析平台队列比较流程

科研人员首先选定总体样本，然后选择腹腔镜手术组和开腹组队列，在设定需观察的变量后即可选择分析方法，最后进行分析，得到分析结果。

4.4 研究结果

系统可以生成定量变量（如年龄和费用）的均数和标准差，同时也能生成两组人群的定性变量的构成比；系统的卡方分析统计结果 OR = 0.017（95% CI 0.003 – 0.039），提示腹腔镜子宫切除术发生并发症的危险性降低；T 检验分析结果，见图2，提示传统腹式组手术费用、药物费用低于腹腔镜组，差异具有统计学意义（$p < 0.05$）。

变量	方差相等	Z值	自由度	P值
手术费用	true	-1.322	441.000	0.037
人均门诊费用	true	0.730	471.000	0.466
人均药物费用	true	-0.697	471.000	0.023
人均住院费用	true	2.170	471.000	0.030

图2 T 检验分析结果

5 讨论和展望

5.1 通用数据模型的意义

OMOP 通用数据模型已累积来自美国、加拿大、澳大利亚、英国等几十个国家和地区的上百个组织机构、超过6 亿人口的临床数据规模，协作研究发表上百篇论文[7]，但是该数据模型在国内的使用仍然较少，OHDSI 专门成立中国工作组，对其提供的产品及服务在中国进行推广，其中包括 OMOP 通用数据模型、中文术语，术语映射工具等[8]。如果国内有更多的医疗机构能够加入，则对国内不同区域甚至全国范围内的医疗大数据进行综合分析的难度将大大降低，有利于进行更大规模的数据分析[9]，从而将分析结果应用于区域或全国的医疗卫生政策制定和临床医疗工作指导，进一步推动整个医疗行业的发展。

5.2 科研分析平台的意义

Vinci 医疗数据科研分析平台将以往耗时费力的医学科研工作简化成可以在网页中迅速高效完成的工作，医生从产生科研设想、获取数据、添加分析变量、设置分析方法到最终得到科研分析结果，时间周期从传统方式的数月缩短为数周，极大地提高科研工作的效率。目前 Vinci 科研分析平台在医院得到广泛应用，如果可以在高校或科研院所等有较多科研任务的单位进行推广，则能够在教育和科研领域发挥更大的作用。一方面促进医学高校教学信息化的发展；另一方面能够在更大范围内提高医疗科研产出质量和效率。

5.3 目前 Vinci 平台中存在的问题

由于目前大多数 OMOP 数据模型中只是抽取医院已有的临床数据，所以 Vinci 医疗科研分析平台目前只能进行回顾性医学研究。如果能够将随访数据集成到 OMOP 通用数据模型，在保证患者隐私的基础上增加数据种类，从而使平台支持前瞻性科学研究是需要解决的问题。

6 结语

本文介绍了医疗临床数据清洗、科研数据中心的建立和Vinci医疗数据科研分析平台的建成及使用，通过实际临床病例应用证明该平台在临床数据分析上的有效性和优势性。Vinci科研分析平台的建立不仅为科研工作者带来便利，也对国家医疗卫生信息系统的发展具有重要意义。此外该数据分析平台也存在着不足，需要在日后的工作中不断完善和改进，使其能够更好地服务于医学科研工作。

参考文献

1 Silverman SL. From Randomized Controlled Trials to Observational Studies [J]. Am J Med, 2009, 122 (2): 114-120.

2 George Hripcsak, Patrick B. Ryanc, Jon D. Dukec, et al. Characterizing Treatment Pathways at Scale Using the OHDSI Network [J]. PNAS, 2016, 113 (27): 7329-7336.

3 OHDSI. org. OMOP Common Data Model [EB/OL]. [2017-07-18]. https://www.ohdsi.org/data-standardization/the-common-data-model/.

4 OHDSI. org. Observational Medical Outcomes Partnership [EB/OL]. [2017-07-18]. http://omop.org/.

5 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.

6 OHDSI. org. Observational Health Data Sciences and Informatics [EB/OL]. [2017-07-18]. https://www.ohdsi.org/.

7 OHDSI. org. OHDSI China [EB/OL]. [2018-01-12]. http://ohdsichina.org/.

8 OHDSI. org. Software [EB/OL]. [2017-07-20]. https://www.ohdsi.org/analytic-tools/.

9 Cithub. Observational Health Data Sciences and Informatics [EB/OL]. [2018-01-12]. https://github.com/OHDSI.