基于通用数据模型的健康医疗大数据平台数据治理研究*

张弘政 刘迷迷 李 琳 承垠林 周 毅

（中山大学中山医学院 广州 510080）

〔摘要〕 以健康医疗大数据平台建设过程中的数据治理实践为例，从数据抽取与清洗、文本数据结构化和数据映射等方面探讨基于通用数据模型的多中心健康医疗大数据质量提升方法和技术，总结相关实践问题与经验，为跨机构、跨部门的健康医疗大数据治理提供参考。

〔关键词〕 健康医疗；大数据平台；通用数据模型；数据治理；数据质量

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.06.001

Study on Data Governance of the Healthcare Big Data Platform Based on Common Data Model *ZHANG Hongzheng, LIU Mimi, LI Lin, CHENG Yinlin, ZHOU Yi, Zhongshan School of Medicine, Sun Yat－sen University, Guangzhou 510080, China*

〔**Abstract**〕 Taking the data governance practice in the construction of the healthcare big data platform as an example, the quality improvement methods and technologies of multi－center healthcare big data based on Common Data Model (CDM) are discussed from the aspects of data extraction and cleaning, text data structuring and data mapping, etc., and relevant practical problems and experiences are summarized, so as to provide references for interagency and cross－sectoral healthcare big data governance.

〔**Keywords**〕 health care; big data platform; Common Data Model (CDM); data governance; data quality

1 引言

1.1 研究背景

随着“互联网＋”、大数据、人工智能、云计算等新兴技术的不断发展和应用，医疗卫生领域信息化程度和水平不断提升，随之产生的健康医疗数据也呈现快速增长[1－2]。这些健康医疗数据多源、多模态、异构且分散存储在不同医疗机构，具有巨大潜在价值，需要以真实世界多中心研究模式统一管理、高效共享和挖掘利用。但是目前我国医疗机构的健康医疗数据存在质量不高[3]、缺乏统一标准[4]等问题，开展多中心的大数据研究困难重重，真实世界健康医疗大数据也难以被真正挖掘和利

〔收稿日期〕 2022－05－23

〔作者简介〕 张弘政，硕士研究生；通信作者：周毅，教授，博士生导师。

〔基金项目〕 国家重点研发计划“友好智慧健康宜居环境系统集成研究”（项目编号：2021YFC2009402）；国家自然科学基金项目“基于非线性动力学驱动的癫痫发作预测深度学习研究”（项目编号：61876194）；广东省自然科学基金项目“基于机器学习的癫痫发作预测脑电及多模态数据模型研究”（项目编号：2021A1515011897）；广东省科技创新战略专项“泌尿系常见病智能诊疗与健康管理关键技术研究与示范”（项目编号：202011020004）；广东省重大科技专项“膀胱癌人工智能一体化精准诊断平台的研发”（项目编号：2018B010109006）。

用，因此亟需开展多中心数据治理，提高真实世界研究数据质量[5-6]。

1.2 政策支持

2018 年中国胸部肿瘤研究协作组发布《中国真实世界研究指南》，对数据源质量、数据采集方案设计、数据标准化等方面提出要求和指导意见，表明了对数据质量控制的重视；2020 年国家药品监督管理局医疗器械技术审评中心发布《真实世界证据支持药物研发与审评的指导原则（试行）》，国家药品监督管理局药品审评中心发布《真实世界研究支持儿童药物研发与审评的技术指导原则（试行）》，表明了临床研究中数据质量的重要性；2021 年国家药品监督管理局药品审评中心发布《用于产生真实世界证据的真实世界数据指导原则（试行）》，从治理、标准和质量保障等方面对数据治理提出具体要求和指导性建议。真实世界研究相关政策文件的相继发布，表明利用真实世界健康医疗数据开展研究成为我国重点发展领域，其中基于数据治理的数据质量提升受到重视。

1.3 研究内容

在此背景下本文面向健康医疗领域真实世界多中心研究，基于通用数据模型相关理论、方法与技术开展健康医疗大数据治理并建立相关研究平台，包括具体实践过程，提高多中心健康医疗大数据质量的关键技术、面临问题与挑战以及解决方案等。经数据治理研究建立的健康医疗大数据平台及相关成果，可为跨机构、跨部门的真实世界研究提供高质量数据，为多中心健康医疗大数据治理提供经验和参考。

2 数据治理与通用数据模型

2.1 数据治理

2.1.1 定义 数据治理是数据资源及其应用过程中相关管控活动、绩效和风险管理等活动的集合[7-8]，具体包括数据标准化、数据质量提升、数据管理和数据应用[9]。数据治理是一个体系性、系统性的集合，不仅通过数据管理提升数据质量，更强调流程设定和权责划分。

2.1.2 内容 目前健康医疗领域多中心真实世界数据治理目的是获得高质量数据用于分析挖掘，提升结论的真实性、可靠性，主要涉及数据标准化和数据质量提升，数据管理和数据应用还有待进一步发展。其中健康医疗数据标准化是参照公认的标准规范，约束健康医疗数据的表达，医务或研究人员按照标准规范记录和使用数据，包括数据抽取与清洗、数据结构化、术语映射等数据规范化以及基于医学信息标准的数据交换和数据集成等[10-11]。而健康医疗数据质量提升，主要内容是构建全流程数据治理体系[12]，即在健康医疗数据治理过程中完善组织架构，明确权利责任分工，使数据质量管理制度化、规范化，实现对数据的产生、共享、使用、统计全过程质量把控以及日常监测、质控和改进；同时建立多中心级的数据标准、含义，梳理分散在不同中心各系统中的数据，参考标准数据集确定统一的命名、定义、数据类型、值域规则、计算方法等。

2.2 通用数据模型

2.2.1 概述 通用数据模型（Common Data Model，CDM）是数据标准化的核心[13]，是具有统一标准的数据模型，可规范健康医疗数据的格式和内容，目的是将不同数据库包含的数据转换为通用格式以及应用统一术语[14]。通用数据模型包含标准化词汇表、标准化元数据、标准化临床数据表、标准化健康系统数据表、标准化健康经济表和标准化派生元素 6 类，共 39 张表，见图 1。

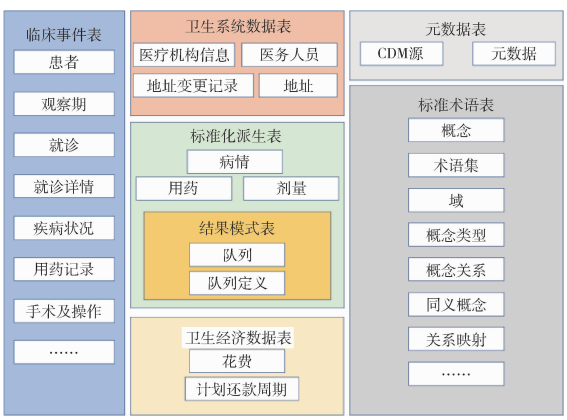

图 1 通用数据模型

2.2.2 健康医疗数据分析和利用标准化 通用数据模型中包含大量医学术语标准并支持开放获取，其中包含世界卫生组织制定的国际疾病分类与代码（ICD10/ICD9）、国际医学术语标准化与研发组织制定的系统化医学术语集临床术语版（Systematized Nomenclature of Medicine—Clinical Terms，SNOMED CT）、美国国立医学图书馆制定的医学主题词表（Medical Subject Headings，MeSH）、观测指标标识符逻辑命名与编码系统（Logical Observation Identifiers Names and Codes，LOINC）、美国国立医学图书馆编制的临床药品规范化命名表（RxNorm）等 100 余个医学术语表。这些术语表为健康医疗数据的分析和利用提供标准化映射的术语支持，健康医疗数据依据统一编码体系和转化规则被标准化为一致概念，基于统一术语表达，后续可开展数据互联互通，检索获取不同医疗卫生机构的数据分析与利用，为大数据研究提供支撑。

2.2.3 健康医疗数据标准化存储 通用数据模型很好地解决健康医疗数据标准化存储问题。通用数据模型具有统一的医学概念表达形式，标准化的临床数据模型、医学术语、编码系统等，数据库内字段信息等属性相对固定。在开展基于通用数据模型的多中心研究时不必考虑适配不同数据库，减少人力、时间投入；通过数据标准指导收集和录入数据，规范了数据采集和管理过程，提高了数据完整性和一致性，保证了研究数据质量。

2.2.4 数据利用 经过基于通用数据模型的健康医疗数据治理，不同医疗卫生机构的信息系统中的健康医疗数据以相同格式的数据结构存储，研究人员可以通过统一的调用方式调取、统计、分析数据，可实现真实世界健康医疗大数据的最大化利用。

3 数据治理实践

3.1 健康医疗大数据平台总体设计

为开展真实世界多中心健康医疗大数据研究，需要对各中心健康医疗数据进行治理，并建立健康医疗大数据平台。在平台的数据治理实践过程中，以通用数据模型为基础建立一套数据入库、清洗、质量检查、结构化、数据映射的标准化处理流程。平台从各个数据中心的不同信息系统中获取患者基本信息、就诊、诊断、用药、检验、手术、文本信息等数据，并进行数据加密与脱敏。取得的数据包括结构化数据和文本数据，对结构化数据直接进行抽取与清洗、质量检查，而对文本数据则利用自然语言处理技术进行实体识别和关系抽取，转化为结构化数据。在对结构化数据和文本数据预处理完成后，针对诊断、手术、药品、检验等数据参照标准术语集分别制定术语映射标准化作业程序（Standard Operation Procedure，SOP），并由医学专家对映射数据进行审校，映射合格的数据即为通用数据模型数据。在这种通用数据模型规范化和标准化的数据基础上开展多中心的临床科研、辅助诊疗、健康管理、疾病预测等应用。

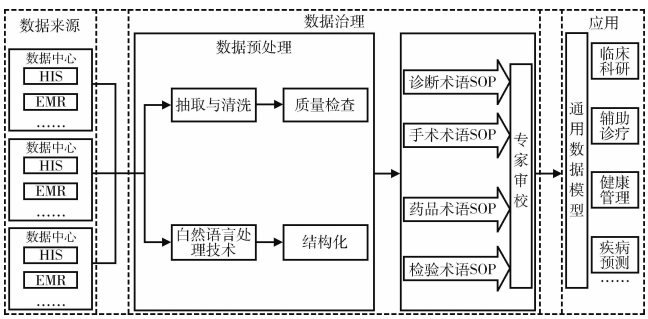

图 2 平台数据治理总体设计

3.2 抽取与清洗

3.2.1 数据抽取 该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，从而确保科研的客观性和患者隐私的保密性。平台支持以接口与非接口化的形式采集数据，支持标准消息传递协议，具备数据存储和访问功能，可将多源异构数据进行聚合。

3.2.2 数据清洗 即对数据中存在的各种问题进行处理，包括缺失值及异常值处理等。处理缺失值时通过统计内容为空、内容缺失数据词频占比确定缺失值数量并进行对应补充；处理异常值时针对部分数据开头或结尾包含特殊字符的情况进行处理，替换掉特殊符号。数据清洗是为了达到补全数据、剔除重复数据等目的，最大限度地利用各中心已有

临床数据，提供更加全面、准确的健康医疗数据。

3.3 质量检查

在数据抽取与清洗完成后对数据进行质量检查，包括完整性检查、关联性检查和一致性检查。完整性检查是将原始数据接口、中间表、通用数据模型的数据量、分布情况进行总体统计检查。一致性检查同样针对以上信息，检查原始数据接口、中间表、通用数据模型库的医疗数据，确保数据一致性。一致性检查要求 100% 一致，对于校验中发现的不一致信息进行评估，检查并更新数据抽取工具或校验工具中的算法。关联性检查对患者基本信息与就诊信息进行关联性检查，并对患者每次就诊信息，包括诊断、用药、检验、手术、文本信息等进行关联性检查，确保患者信息与业务数据是准确关联的。关联性检查的要求是患者基本信息、就诊信息与业务场景相符合。

3.4 数据结构化

除了结构化数据，医疗数据还包括大量非结构化文本数据，需要利用自然语言处理技术抽取这些文本数据的实体和关系进行结构化。首先通过机器学习构建命名实体识别（Named Entity Recognition，NER）模型和关系抽取模型，提取文本中的实体和关系。其中实体指的是文本中的信息字段，可分类为疾病诊断、时间、药品名、症状表现、值等，关系是指两个或多个实体之间存在的逻辑关系。文本数据结构化后进行校验，针对命名实体识别模型和关系抽取模型在实体和关系提取时的可信度（即模型的准确率）方面进行验证，确保模型的准确性达到 90% 以上，主要用准确率、召回率、*F*1 值 3 个指标衡量文本数据结构化处理效果，3 个衡量指标均≥90% 则可认为文本数据结构化处理质量达到要求。

3.5 术语映射

3.5.1 概述 平台数据治理中的术语映射为半自动化术语映射，即机器为主、人工为辅。标准概念由机器学习算法自动推荐，并由医学人员逐条确认映射结果；对不同类型的术语制定术语映射标准作业程序，保证术语映射规则统一。平台的术语映射主要包括数据质量评估、数据拆分、术语映射、专家审校等内容。

3.5.2 数据质量评估 包括评估数据类型，分析待映射数据包含的数据类型、种类，确定待映射术语体系；评估数据完整性，分析数据是否存在缺失值及异常值等，针对存在问题及时反馈；评估整体情况，分析是否需要处理缺失值及异常值，以及是否有分词需求。

3.5.3 数据拆分 对不同类型数据参考不同术语体系做标准化，因此对包含多种类型的源数据按类别拆分后再进行映射，见表 1。

表 1 不同数据类型对应的术语集

数据类型	术语集
诊断数据	标准术语集
检验数据	LOINC
药品数据	NCCD
手术数据	ICD9CM

注：标准术语集为中国已发布的标准术语集；ICD9CM 为第 9 修订版的疾病临床修正国际分类。

3.5.4 术语映射 分为自动映射和人工映射。自动映射自动匹配标准术语库，完全精准匹配结果不再进行人工映射；人工映射时，首先基于机器学习算法自动推荐标准概念，并给出一个基于概率的置信度（0～100 之间），然后通过机器自动推荐结合医学人员手动搜索对照标准的术语字典逐条映射，并标注映射状态为“近似精准”“向上映射”“存疑映射”或“无法映射”，见表 2。

表 2 术语映射规则

术语分类	映射状态	状态定义
诊断术语映射规则	完全精准映射	待映射术语与标准概念一致
	近似精准映射	待映射术语与标准概念语义一致，表述不同

续表 2

	向上映射	无法精准映射情况下，向上映射至较为宽泛的上位概念
	存疑映射	无法确定映射准确性
	无法映射	无法精准或向上映射
手术术语映射规则	完全精准映射	待映射术语与 ICD9CM 标准概念一致
	近似精准映射	待映射术语与 ICD9CM 标准概念语义一致，表述不同
	向上映射	无法精准映射情况下，向上映射至较为宽泛的上位概念
	存疑映射	无法确定映射准确性
	无法映射	无法精准或向上映射
药品术语映射规则	精准映射	待映射术语与 NCCD 标准概念一致
	向上映射	10ml：10g，如需计算，则映射 1g/1ml
		无法精准映射情况下，依次按照化学名、剂型、剂量、商品名顺序向上映射至 NCCD 标准概念
	存疑映射	待映射术语与 NCCD 标准概念语义一致，表述不同
	无法映射	无法精准或向上映射
检验术语映射规则	精准映射	名称、样本、单位类型一致
	近似精准映射	名称、样本一致，单位可换算
	向上映射	名称、单位一致，样本可向上映射对应的 LOINC 概念
	无法映射	无法精准或向上映射

3.5.5 *专家审校* 由医学专家审校映射准确性，针对近似精准、向上映射、存疑数据由医学专家对照标准集进行校验及更正，针对无法映射术语由专家再次映射，仍无法映射的术语由专家指导构建标准术语并补充到标准术语集，最终更新至映射规则库。

3.5.6 *质量核查* 术语映射质量核查即随机抽取 10% 的映射数据，如果映射准确性≥90% 则认为数据映射合格。

3.6 阶段成果

3.6.1 *概况* 平台经基于通用数据模型的数据治理汇聚 3 个医疗机构的健康医疗数据，包括 131 万患者数据，其中住院患者数据约 12 万，门诊患者数据约 117 万，手术患者数据 9 万，检查检验数据约 3 000 万。平台具有数据概览、探索发现、队列发现、科研管理等功能模块，能够支持科研人员高效、便捷地研究、统计、管理和分析患者数据，提高研究效率，拓展研究范围。

3.6.2 *数据概览方面* 支持对平台全量数据及建立特定队列的患者数量、住院患者数量、门诊患者数量、手术数量、检查检验数量、性别、年龄、地域分布等数据进行统计与可视化，对数据进行描述性统计并以多种图表的方式呈现，使研究人员快速了解数据总体情况。

3.6.3 *数据检索方面* 对通用数据模型的健康医疗数据建立索引，通过搜索引擎快速、准确地搜索相关结果并排序。支持通过常用信息、病案首页、检查信息、治疗信息、用药信息、检验信息等进行检索。其中检查信息包括影像检查和检查基本信息，治疗信息包括手术信息，检验信息包括基本信息和常用检验项目。以此自定义条件检索出符合条件的人群进行探索性分析和队列发现，同时支持建立队列，自动汇聚和采集满足队列纳入排除标准的回顾性数据和前瞻性数据，并支持合并多个研究队列。

3.6.4 *科研管理与数据分析方面* 支持前瞻性和回顾性的科研项目建立、查看、资料修改与完善、数据使用、数据导出等科研管理。同时平台集成了 *T* 检验、卡方分析、方差分析等常用卫生统计方法，支持简单的数据分析与统计；支持将队列筛选和变量选择所得数据导出，在更专业的统计分析工具中开展更深入的数据分析和挖掘。

4 存在的问题与建议

4.1 概述

在平台的数据治理实践中实现了真实世界多中

心健康医疗数据标准化和质量提升。通过制定不同的数据治理标准作业程序，将不同医疗机构质量参差不齐、结构各异的健康医疗数据转换成通用数据模型格式，为真实世界多中心健康医疗研究提供高质量、高可靠的支撑。但在数据治理过程中还存在一些问题，也是当前真实世界多中心健康医疗大数据治理研究的普遍问题。

4.2 信息系统维护不足

医疗机构以满足临床业务为主，对信息系统的维护不足，使得信息系统里的数据存在不完整、不规范、不标准、缺乏关联等问题。医疗机构应将数据作为资产管理，以通用数据模型为基础构建全流程的数据治理体系，做好数据日常维护，以减少多中心研究中在单中心数据质量控制上的人力、物力消耗。

4.3 模型实体识别和关系抽取能力有待提升

非结构化中文文本数据存在歧义性和记录信息不完整等问题，加上医疗概念复杂，自然语言处理模型难以处理医学领域的常识和推理问题。可尝试以病种为单位划分数据和搭建单病种知识图谱，以点带面构建行业内的常识性知识，并进一步开展受限自然语言处理，提升模型的实体识别和关系抽取能力。

4.4 数据不够全面

由于目前平台集成了各中心一部分健康医疗数据，数据还不够全面，对研究结果可能有一定影响，需要补充影像、基因、随访等更多模态、来源的数据，同时需要保证数据安全和患者隐私。为此可尝试在平台上接入更多基于通用数据模型的数据处理、分析与挖掘的统一代码或工具；各中心利用分布式网络调用平台提供的代码或工具对医疗数据进行治理，存储在本地，并对数据进行分析和挖掘，共享研究结果。各中心不需要输出可能包含患者隐私的数据，只需要将研究结果整合起来，对外仅分享和发布整合研究结果。

4.5 多中心健康医疗大数据应用发展相关法规政策缺位

目前我国未出台专门针对多中心健康医疗大数据应用发展的专项法律法规、配套政策及监督机制等，存在数据的归属权与使用权不明确、数据共享开放的管理制度以及应用准入与退出机制缺乏、数据应用的公平性机制不清晰等问题，制约了我国健康医疗大数据的良性发展。因此需要在国家层面对规范数据质量、数据来源的合法性、数据采集的合规性、个人信息授权、数据脱敏化处理、数据应用的公平性等一系列健康医疗大数据应用过程中的环节制定详细的政策法规和体制机制。

5 结语

本文通过建立真实世界多中心健康医疗大数据平台，提供基于通用数据模型、统一的理论、方法与技术，实现多中心健康医疗数据的规范化和标准化，提高数据质量，推动跨机构、跨部门的数据互联互通和共享利用，使真实世界健康医疗大数据真正成为资源，发挥应有价值。

参考文献

1 郭强，王丛，衡反修. 医疗大数据平台建设机遇、挑战及其发展［J］. 医学信息学杂志，2021，42（1）：2－8.

2 王觅也，郑涛，李楠，等. 医疗大数据集成及应用平台体系构建［J］. 医学信息学杂志，2019，40（8）：37－42.

3 姬卫东，李琳，张振，等. 互联互通背景下医疗数据治理面临的问题与对策［J］. 中国数字医学，2021，16（11）：6－11.

4 龙思哲. 基于数据中台的医院信息系统数据治理方案的探讨［J］. 当代医学，2021，27（29）：193－194.

5 王强，易应萍. 临床医疗大数据治理和应用［J］. 医学信息学杂志，2018，39（8）：2－6.

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

7 邓军增. 医院健康医疗数据治理探讨［J］. 医学信息学杂志，2021，42（8）：14－17.

8 国家市场监督管理总局，中国国家标准化管理委员会. GB/T 34960.5－2018 信息技术服务 治理 第5部分：数据治理规范［EB/OL］.［2021－06－03］. https://max.book118.com/html/2019/0918/5031002343002130.shtm.

（下转第 13 页）

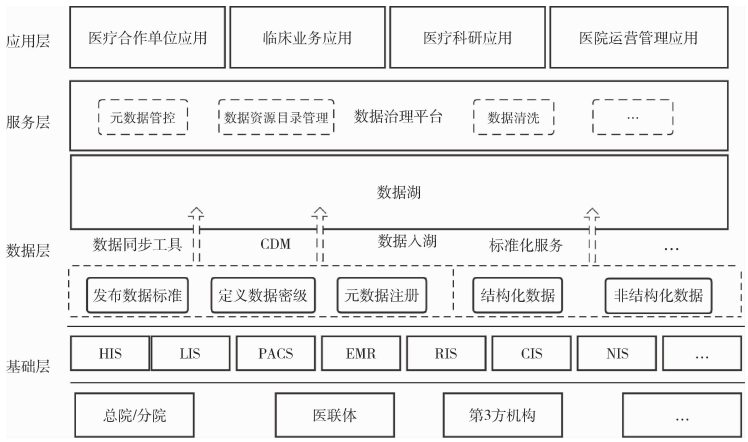

图 3 数据湖建设框架

6 结语

目前借助数据湖所要达成的目标涉及不止一种数据技术，汇集了包括数据仓库、实时和高速数据流技术、数据挖掘、深度学习、分布式存储等技术在内的多种技术，已经从一种“大数据存算方案”进阶到“大数据存算 + 处理分析 + 资产治理 + 安全隐私 + 数据变现”一揽子方案。在数字经济时代，从数据仓库到数据湖不仅是数据存储架构的变革，更是大数据思维方式的升级。数据湖能为医院赋能，帮助医院优化运营模型，为医院科研提供更多维度数据分析，有助于医院提升运营管理和科研能力。

参考文献

1 王韶锋，赵善斌，杨静. 医院数据治理与数据质量提升研究［J］. 现代医院，2021，21（11）：1761 – 1763.

2 尹西明，林镇阳，陈劲，等. 数字基础设施赋能区域创新发展的过程机制研究——基于城市数据湖的案例研究［EB/OL］.［2022 – 06 – 07］. http://kns.cnki.net/kcms/detail/12.1117.g3.20220509.1258.002.html.

3 李硕，卢华明. 基于数据湖的环境大数据存储模型［J］. 北京信息科技大学学报（自然科学版），2021，36（6）：81 – 86.

4 任仲晟. 基于数据仓库的数据挖掘技术［J］. 数字技术与应用，2021，39（9）：59 – 61.

5 华为数据管理部. 华为数据之道［M］. 北京：机械工业出版社，2020.

6 王志勇，吴骋，王立鹏，等. 医疗大数据背景下的数据治理与质量监管［J］. 中国数字医学，2021，16（4）：92 – 96.

7 叶琳，罗铁清. 医疗数据治理综述［J］. 计算机时代，2021（5）：10 – 12.

8 王志勇，吴骋，王立鹏，等. 医疗大数据背景下的数据治理与质量监管［J］. 中国数字医学，2021，16（4）：92 – 96.

9 吴信东，董丙冰，堵新政，等. 数据治理技术［J］. 软件学报，2019，30（9）：2830 – 2856.

10 邓军增. 医院健康医疗数据治理探讨［J］. 医学信息学杂志，2021，42（8）：14 – 17.

（上接第 7 页）

9 龙思哲. 基于数据中台的医院信息系统数据治理方案的探讨［J］. 当代医学，2021，27（29）：193 – 194.

10 叶琳，罗铁清. 医疗数据治理综述［J］. 计算机时代，2021（5）：10 – 12.

11 董方杰，李岳峰，杨龙频，等. 我国卫生健康信息标准工作进展与展望［J］. 中国卫生信息管理杂志，2019，16（4）：400 – 405.

12 王韶锋，赵善斌，杨静. 医院数据治理与数据质量提升研究［J］. 现代医院，2021，21（11）：1761 – 1763.

13 Wang Q，Reps J M，Kostka K F，et al. Development and Validation of a Prognostic Model Predicting Symptomatic Hemorrhagic Transformation in Acute Ischemic Stroke at Scale in the OHDSI Network［J］. PLoS ONE，2020，15（1）：e0226718.

14 洪娜，刘飞，张梦阳，等. OHDSI 通用数据模型在肿瘤大数据中的应用探索［J］. 中国数字医学，2021，16（11）：24 – 28.