基于全民健康信息平台的医疗健康大数据治理方法及技术研究

徐 静 高昭昇 黄岳源 吴宇婷

（广州市卫生健康技术鉴定和人才评价中心 广州 510080） （卫健智能（深圳）有限公司 深圳 518000） （广州市卫生健康技术鉴定和人才评价中心 广州 510080）

〔摘要〕 基于全民健康信息平台多年来采集的海量数据，通过低质量数据过滤、患者主索引匹配、医学术语映射、医学文档结构化处理、医疗数据隐私化处理等大数据治理方法及技术，实现对医疗健康数据质量评价与控制，提高医疗健康大数据应用分析价值，支撑医疗健康大数据持续性的创新应用。

〔关键词〕 大数据治理；openEHR；医学术语映射；数据质量评价

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.07.002

Study on Medical and Health Big Data Governance Method and Technology Based on National Health Information Platform

XU Jing, GAO Zhaosheng, Guangzhou Health Technology Appraisal and Talent Evaluation Center, Guangzhou 510080, China; HUANG Yueyuan, Medical Health Intelligent (Shenzhen) Co. Ltd., Shenzhen 518000, China; WU Yuting, Guangzhou Health Technology Appraisal and Talent Evaluation Center, Guangzhou 510080, China

〔**Abstract**〕 Based on the massive data collected by the National Health Information Platform over the years, through big data governance methods and technologies including low－quality data filtering, Enterprise Master Patient Index (EMPI) matching, medical term mapping, medical document structured processing, medical data privacy processing, etc., the quality evaluation and control of medical and health data is realized, the application and analysis value of medical and health big data is improved, and the sustainable innovative application of medical and health big data is supported.

〔**Keywords**〕 data governance; openEHR; medical term mapping; data quality assessment

1 引言

自 2009 年起全国各地陆续启动以居民电子健康档案为核心的全民健康信息平台建设。平台主要采集居民在医疗卫生机构的诊疗信息，完整记录居民全生命周期健康信息及其他相关信息。随着平台覆盖范围进一步扩大，基于平台的应用不断深化，积累了海量临床诊疗、检验检查、医院运营等数据[1]。然而质量问题成为大数据应用面临的挑战，高质量的数据才能更好地支撑各类应用；同时对于个性化健康指导、基于群体的卫生经济负担研究而言，数据质量直接影响研究结果的准确性、科学性。由于数据在产生和存储过程中质量较难控制，因此在采集过程中进行评估以提前了解数据质量对于后续数据应用

〔收稿日期〕 2022－05－11

〔作者简介〕 徐静，硕士，工程师，发表论文 12 篇，参编著作 5 部；通信作者：高昭昇，博士，高级工程师。

或质量提升具有重要意义。通过持续推进统一权威、互联互通的全民健康信息平台建设，基于区域平台实现横向、纵向数据共享已形成广泛共识，随着卫生健康事业的发展、新政策的实施，平台被赋予了更多更高的要求，如实现全面智慧监管、助力健康产业发展等。通过数据治理能逐步完善平台并推动其高质量发展。本文对大数据治理方法及技术进行探索研究，以期为医疗健康大数治理及有效利用提供参考。

2 现状及存在的问题

2.1 现状

我国已有 31 个省份完成全民健康信息平台建设，平台中汇聚了医疗服务、公共卫生、医疗保障、医院运营等数据，大部分来源于各级医疗卫生机构，与公安和医保部门数据共享比例较高[2]。通过全民健康信息平台建设，管理者可以及时了解医疗卫生服务资源使用情况，便于实施高效、精准决策，进而有效减少医疗成本，加强医疗服务质量和提升医疗服务效率；医护工作者可以随时随地获取所需信息，实现高质量、高效率医疗服务；公共卫生工作者可以掌握居民健康信息，便于提前开展疾病防控和健康促进相关工作；居民可以获取并全面了解自身健康信息，有助于实现个人健康管理以及享受及时、便捷、持续的医疗服务。

2.2 存在的问题

医疗健康大数据来源于不同医疗卫生机构、信息系统，通过全民健康信息平台可以实现跨区域医疗数据整合汇聚，但尚存在数据应用效果不佳、未能真正发挥数据价值等问题，主要表现在以下几方面：一是整合后的数据质量参差不齐，表数据缺失、表关键字段缺失等数据缺失率严重，数据种类不符、乱码、索引号混乱等现象频发，医学术语不统一，存在大量计算机很难识别的非结构化文本、影像、视频等，数据应用价值不高，档案调阅量低。二是医疗机构信息化建设水平决定其整合数据的数量与质量，大量健康信息缺乏信息系统记录，为个体疾病、健康状况信息收集和分析带来困难[2]。三是数据缺少有效性、正确性校验，数据核查困难，安全性需要有效提升，各医疗机构数据质量参差不齐，导致针对医院的评价结果准确性不高，较难实施有效监管及绩效评价。

3 医疗健康大数据治理整体架构设计

3.1 概述

数据治理的最终目标是提升数据价值，通过医疗健康大数据治理实现数据定义、模型构建、数据质量评估、数据应用等环节的统一。基于数据定义标准，对业务指标应用进行结构化拆解，实现指标的技术定义，完成模型设计；基于模型设计环节沉淀的元数据，驱动和约束最终物理模型设计，为数据加工确定最终数据定义语言（Data Definition Language，DDL），完成物理模型设计，以此约束后续数据开发。

3.2 整体架构

区域全民健康信息平台承担健康数据采集、汇总与应用功能，主要包括：一是将数据上传到信息资源中心，包括前置机采集数据；二是对已上传数据进行合理性校验，并将其存储至信息资源中心，按照数据自身特点进行相应转换、管理和分发；三是满足各方对信息资源中心临床数据调阅等应用需要。数据治理首先要研究医疗健康大数据有效利用的精细化信息模型，然后通过低质量数据过滤、患者主索引匹配、医学术语映射、医学文档结构化处理、医疗数据隐私化处理等治理技术对符合信息模型的数据进行半自动化处理，实现较高的自动处理准确率。通过平台汇聚、评估、治理的不断迭代过程，建成可快速检索的高质量医疗大健康大数据资源，支撑医疗健康大数据持续性创新应用研发，实现包括疾病预警预测、临床决策支持、个人健康画像、高危筛查等大数据应用，见图 1。

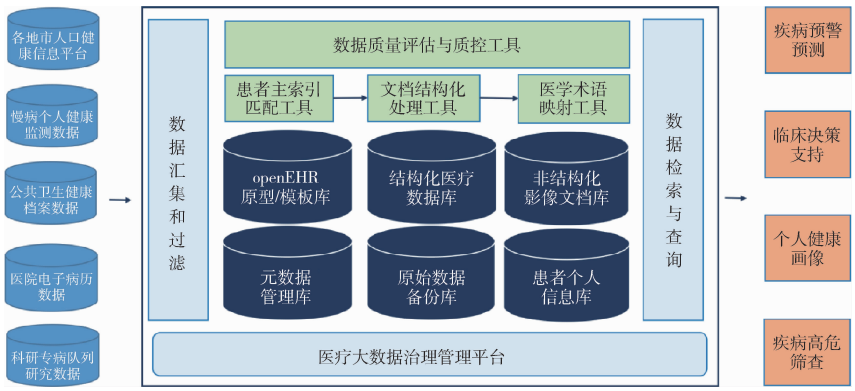

图 1 医疗健康大数据治理整体架构

3.3 数据模型构建

3.3.1 *构建流程* 对应用场景数据需求进行详细剖析，包括医疗服务（门诊 + 住院）、公共卫生、中医、专病分析等领域的应用场景，基于数据需求详细剖析结果，参照国家或国际对于数据元相关标准，定义标准化数据元，通过元数据标准定义的元数据项对数据集进行描述[2-3]。调研国际各类医学术语标准，参照国际对于数据值域的相关标准，针对每个数据元选定最合适的医学术语或值域。对 OpenEHR 原型进行扩展，建立精细化的 OpenEHR 原型库并根据实际数据源定义精细化的 OpenEHR 存储模板，用于指导大数据中心结构化数据库构建。对多来源的各类数据进行汇集，建立各数据源的元数据并进行统一管理，包括但不限于全民健康信息平台项目数据、个人持续健康监测数据、医院电子病历数据、公共卫生信息平台数据及专病队列研究数据等。

3.3.2 *模型构建* 通过对患者处方、检验报告、检查报告、体检报告等医疗健康大数据进行规则和类型定义，根据不同数据特性建立多个质控模型。通过 OpenEHR 原型模板编辑工具，针对个人居民健康档案基本信息、疾病控制管理、儿童保健、妇女保健、医疗服务等关键业务域，对每个关键业务域的附属数据按照其所在业务进行分类，将关键业务域与附属业务数据按照数据关键字进行关联，再对所有关键业务域按照个人主索引进行关联，从而形成个人居民健康档案数据模型。

3.4 质量评估与控制

3.4.1 *数据质量评估框架模型* 基于国际数据质量标准，结合国内医疗领域数据质量标准进行细化扩展，建立详细的数据质量评估框架模型，从数据完整性、一致性、规范性、结构化、逻辑性等方面指导分析数据质量评估需求以及制定数据质量评估规则。

3.4.2 *数据质量规则定义语言与规则编辑管理工具* 基于规则定义语言（Guideline Definition Language，GDL），定义临床数据质量评估规则，对数据质量模型中各种特征的度量标准进行衡量。开发质量规则编辑管理器，实现对质量评估规则的扩展与维护。

3.4.3 *针对临床应用需求构建数据评估规则库* 分析数据应用的实际需求，构建临床数据使用场景的质量评估规则库。根据设定规则自动纠正数据值错误、数据类型错误、半角全角字符、中英字符错误等一般数据质量问题；针对其他复杂数据质量问题，利用专业数据质量提升算法的修复技术加以解决，包括且不限于缺失数据填补、标准术语映射、重复数据检测与删除等。对所采集医疗数据，如处方、检验报告单、检查报告单、体检报告单等进行规则和类型定义，根据不同医疗数据特性建立多个质控模型。

4 医疗健康大数据治理技术分析

4.1 数据处理

4.1.1 *数据错误处理* 数据中心数据转换、加载

过程的第 1 步。数据经过接口初步验证后开始进行错误处理。主要是检查数据源数据格式是否满足相关标准，内容是否正确。数据源可分为两类：数据文件和数据库接口数据。数据源的数据需要校验和转换后才能进入数据中心，见图 2。

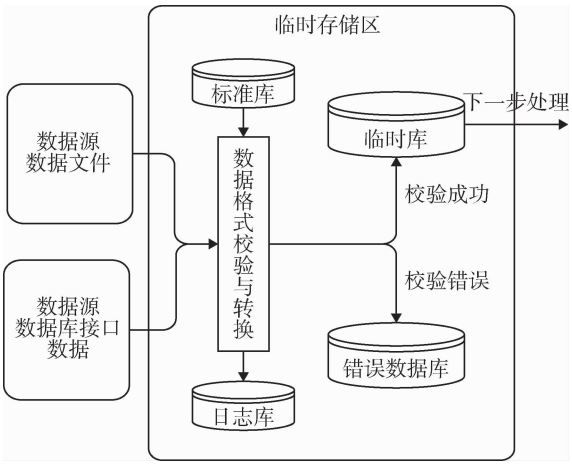

图 2 数据错误处理

4.1.2 数据清洗 在数据临时存储区中进行，是为保证数据质量而进行的必要的数据清理工作，其目的在于保证数据质量和准确性。由于卫生健康数据的多样化，要保持其唯一性和可靠性需要进行数据清洗。数据清洗需设定规则，规则存储在标准库中，在进行清洗过程中调取。一般来说清洗过程是数据提取、转换、加载（Extract – Transform – Load，ETL）过程的一部分。清洗后的数据经过业务处理即可进入操作型数据存储库（Operation Data Store，ODS）中[3-4]，见图 3。

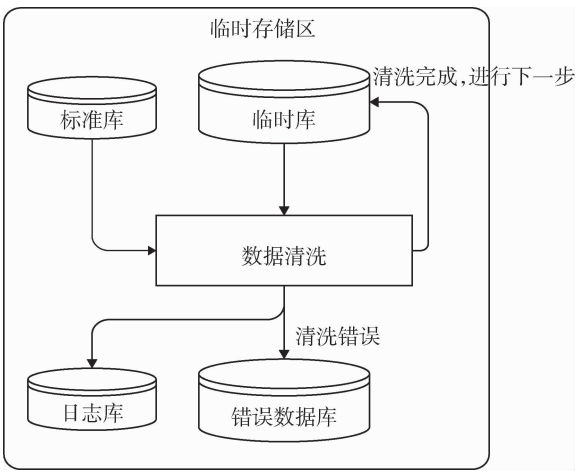

图 3 数据清洗过程

4.1.3 非结构化数据 指不便于使用数据库二维逻辑表呈现的数据，包括所有格式的办公文档、报表、文本、图片、HTML、XML、图像、视频和音频信息等，可以采用跨企业文档共享（Cross – Enterprise Document Sharing，XDS）模型为其建立索引服务，实现非结构化数据的存储与访问。XDS 集成规范定义了基于 XDS 模型的核心基础服务。其他许多信息技术基础设施（Information Technology Infrastructure，ITI）集成规范则在安全、患者标识管理、文本就绪通知和新的文本交换路径等方面进一步加强 XDS 基础设施。

4.1.4 医学术语映射 指在原始数据实际取值和标准数据术语概念之间建立联系，在导入过程中对数据进行基于语义的标准化处理。将原始数据进行规范化操作后需明确使用术语服务的对象字段。通常这类字段内容具有表述简单、意义明确、逻辑性强等特征，例如药物剂型、剂量，物质存在状态等，而不是文字描述性内容。实现相应术语服务步骤如下：第一，确定该字段原始数据取值范围；第二，将原始数据所有取值与该字段对应 HL7 项的词汇域中所有取值一一对应；第三，为第 2 步涉及的所有概念建立别名，别名取值即为相应字段原始数据取值。以“临床用药”数据中的“剂型”字段为例，首先考察原始数据，可知 DrugType 的取值为“口服液”“片剂”“喷雾剂”3 种；其次在由数据规范化确定的 DrugType 对应的 HL7 项中找到与这 3 种取值对应的词汇，形成术语映射表，见表 1。为“Oral Solution”“Tablet”和“Oral Inhalant”3 个概念创建别名，取值分别为“口服液”“片剂”和“喷雾剂”。在数据导入过程中数据的 DrugType 字段自动实现基于语义的标准化。

表 1 “临床用药”DrugType 字段的术语映射

原始数据	术语取值	术语代码	编码方案	编码方案版本
口服液	Oral Solution	S14431	HL7 Vocabulary	v3.0
片剂	Tablet	S14515	同上	同上
喷雾剂	Oral Inhalant	C14565	同上	同上

4.2 数据服务

数据服务主要为数据治理提供统一的数据建

模、数据管理、数据应用等服务，数据中心需要支持各业务主题域、主题分析、数据集市模型和指标相关的语义和解释的定义，该定义包括业务口径和技术口径，并支持数据模型调整。同时在数据访问层或数据服务层面提供数据可视化服务和清晰的数据指标定义，见图 4。

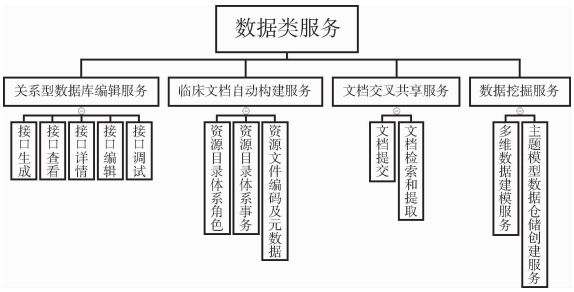

图 4 数据类服务

4.3 安全与隐私治理

在电子健康档案共享过程中，为保障居民隐私安全必须应用数据安全加密处理技术，实现对平台核心采集交换信息加密存储；设置隐私保护服务模块，通过身份验证、短信验证、核心数据字段隐藏等多种方式实现居民健康档案信息调用隐私保护[5]。通过匿名化服务确保在平台中以及提供正常医疗服务以外（例如医疗保险、管理以及某种形式的研究）的资料传递过程中，不会面向非授权用户泄露患者身份信息，将患者身份信息通过移除或修改等手段进行泛化处理。

5 数据治理结果评价

全民健康信息平台在完成数据治理后需要构建治理结果评价体系，对治理结果进行评价或绩效考核，数据治理结果评价应围绕数据质量和数据安全两部分进行，数据质量按照互联互通成熟度测评指标要求，一般分为数据逻辑性（关联性、约束性）、及时性、完整性（一致性）、稳定性、准确性，对数据质量结果进行评分，根据业务要求配置质控规则，包括健康档案业务、综合管理与决策支持业务等。通过质控规则判定数据质量及评分并可及时发布质控问题给医院，协助医院处理质控问题，实现数据全程质量控制，进而达到计划 - 执行 - 检查 - 处理（Plan - Do - Check - Action，PDCA）的流程循环控制[6-7]。完成数据治理后数据质量得到进一步提升，治理结果可以根据平台相关业务考核标准及要求，通过计算机化、通用化处理形成最小单位的规则标准，在规则标准基础上制定考核模板，包括规则类型设定、指标类型设定以及逻辑化和业务标准化，系统最终以规则模板为基础对数据进行全量计算，得出考核结果和考核评分[8]。

6 结语

通过区域全民健康信息平台数据治理，医疗卫生信息可以在区域内医疗卫生服务机构中实现互通共享，便于各级各类医疗卫生机构进行协作，实现全市医疗服务高效率、高质量发展。数据质量提升是一个漫长且复杂的过程，需要根据数据应用需求、应用场景随时调整策略，如果没有建立有效的数据治理机制将影响数据可用性、创新技术应用、平台效率及有效性，需确保数据在用于分析及决策前可靠、一致。持续研究大数据治理方法及技术，实现对医疗健康大数据质量评价与控制，进一步提高应用服务的可用性和效率，通过治理形成高质量的医疗健康大数据资源，是平台建设面临的迫切任务。

参考文献

1 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.
2 《中国卫生信息管理杂志》编辑部. 全民健康信息化调查报告——区域卫生信息化与医院信息化（2018 年）专家审校会在京召开 [J]. 中国卫生信息管理杂志，2019，16（2）：124.
3 许文韵. 健康医疗大数据中心建设实践与思考 [J]. 医学信息学杂志，2020，41（8）：48 - 51，56.
4 俞鹏飞，罗颢文，刘建模，等. 面向医院的大数据治理模型设计 [J]. 医学信息，2021，34（10）：18 - 20.
5 阮彤，邱加辉，张知行，等. 医疗数据治理——构建高质量医疗大数据智能分析数据基础 [J]. 大数据，2019，5（1）：12 - 24.
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 卫荣. 健康医疗大数据质量治理研究 [J]. 中国卫生质量管理，2020，27（3）：5 - 8.
8 孙宗锟. 面向智慧医疗云平台数据使用的隐私保护研究 [D]. 青岛：山东科技大学，2020.