《中国卫生质量管理》第29卷 第3期(总第172期)2022年3月

基于省级健康信息平台的数据采集及治理研究*

——周莉莉 徐 进 孙润康 汪火明*

【摘 要】 基于湖北省级健康信息平台的数据采集及治理方案，分析了从数据源选取、数据采集、数据存储到数据治理全流程的相关情况。认为：数据采集应遵循一定的采集规则和技术要求；数据治理包含数据清洗、数据标准化处理和结构化处理、元数据处理、数据质控、数据脱敏与加密处理等。健康信息平台的应用可为其他省级健康信息平台进行数据采集与治理提供借鉴。

【关键词】 省级平台；健康信息；数据采集；数据治理

中图分类号：R197.324 文献标识码：B

Data Collection and Governance Based on Provincial Health Information Platform/ZHOU Lili, XU Jin, SUN Runkang, et al.//Chinese Health Quality Management, 2022, 29(3): 65－68

Abstract Based on the data collection and governance scheme of Hubei provincial health information platform, the whole process from data source selection, data collection, data storage to data management was analyzed. It was considered that data collection follows certain collection rules and technical requirements; data governance includes data cleaning, data standardization and structured processing, metadata processing, data quality control, data desensitization and encryption processing, etc. It can provide reference for other provincial health information platforms to collect and manage data.

Key words Provincial Platform; Health Information; Data Collection; Data Governance

First-author's address Tongji Hospital, Tongji Medical College of Huazhong University of Science and Technology, Wuhan, Hubei, 430030, China

1 研究背景

2020年6月28日，国家卫生健康委员会办公厅发布《关于做好信息化支撑常态化疫情防控工作的通知》(国卫办规划函〔2020〕506号)，要求强化新冠肺炎疫情信息监测预警、应急指挥与大数据综合分析。在新冠肺炎疫情防控期间，所有省(自治区、直辖市)均要进行卫生事件数据上报，省级数据上报至国家平台，国家平台对数据进行处理，为管理决策提供数据支持。但由于缺乏统一的数据采集标准，上报的数据格式混乱，且结构不清晰。鉴于此，由湖北省卫生健康委员会牵头，建立了湖北省医疗健康大数据平台[1]，其中华中科技大学同济医学院附属同济医院承担了子课题项目，主要研究内容为：省级平台如何向国家平台上报数据、哪些数据需要上报、上报需要遵循怎样的数据格式标准与管理标准等。

本研究以省级平台与国家平台对接为背景，着重分析医疗健康数据采集及治理方案，具有以下意义：(1)可为各省级平台与国家平台对接提供数据采集及治理方案；(2)对加强区域数据上报、实现信息互联互通具有促进作用；(3)为应对突发公共卫生事件提供了数据支撑。

2 调研分析

2.1 数据采集现状分析

2.1.1 业务数据现状分析 目前，

DOI：10.13912/j.cnki.chqm.2022.29.03.18

*基金项目：国家高技术研究发展计划(863计划)——数字化医疗关键技术集成与应用示范(2012AA02A612)

周莉莉 徐 进 孙润康 汪火明* 通信作者：汪火明

华中科技大学同济医学院附属同济医院 湖北 武汉 430030

湖北省卫生健康委员会已建立了湖北省全民健康信息平台，实现了省域内全民健康信息省、市、县三级平台的互联互通。其中，武汉市已建成市人口健康信息平台，各区县也完成区域内人口健康信息的采集及整合，包括居民电子健康档案信息、电子病历信息、人口信息资源库信息以及其他必要的卫生、计生监管信息等。同时，基于全民健康信息平台，湖北省正在筹建医疗健康大数据中心，计划完成数据采集以后向国家平台进行数据上报。

2.1.2　数据采集技术现状分析

目前，省级健康信息平台数据上报格式多样，未实现统一的数据筛选、数据治理、数据整合方式。有的使用文件上传形式实现上报，文件格式包括 excel、cda、csv、xml 等；有的采用数据接口对接实现上报，接口对接方式采用 SOAP、Web Services、Ldap 等；有的采用数据库实现上报，上传数据库格式包括 csv、dat、dbf、mdb、odb 等。另外，文件上传上报形式多样，标准不统一；数据接口对接上报形式缺乏统一的采集接口规范，存在残缺数据；数据库上报形式缺乏统一的技术平台进行数据处理等。这些都是当前数据采集存在的重要问题。

2.2　数据采集问题分析

(1)医疗、疾控等系统数据未进行有效整合，造成区域内卫生、疾控数据管理效能低下[2]。如在新冠肺炎疫情防控期间，与患者相关的密切接触者、核酸检测、物资申领、健康码、出入境、购药等数据无法有效联动。

(2)由于医院临床工作和疾控流调工作分属不同系统，在发生公共卫生事件时医务人员是一线接触患者的工作人员，但却无法获得完整的区域性、时域性疾病数据。与此同时，疾控流调人员也无法获得完整的确诊患者的临床诊疗数据，造成双方协作工作延迟，重复性数据采集问题频发。

(3)医疗数据如不进行有效的数据治理，则无法开展回顾性研究，不能形成有效的研究成果，也就无法形成支撑相关成果转化的源动力。

(4)未进行有效数据治理的医疗数据质量普遍不高，难以直接利用，倘若要联合交通、民政、公安等行业数据进行分析就更加困难。如何对多行业数据进行有效的数据治理及数据整合，是重大公共卫生事件监测体系建设的关键。

3　数据采集方案

数据采集流程涉及数据采集、数据流、数据存储、数据加工与处理以及数据服务、数据分析决策等。本研究基于省级健康信息平台的数据采集流程，其调研数据来源于市级人口健康信息平台等数据库[3-5]。数据被采集以后进入数据湖，与业务数据采集同时进行，不直接从业务数据库中提取，不会对业务数据产生干扰。数据湖中的数据经过数据同步、数据抽取、数据清洗和数据转换[6]等进入数据中台，支撑应用层的数据应用服务[7]。在整个数据采集流程中，通过对数据实时整合、控制消息队列、批量数据抽取、数据转换[8]等，对数据质量进行管控。基于省级健康信息平台的数据采集流程见图 1。

3.1　数据采集范围

数据采集范围包含多个数据源，本研究讨论典型数据源[9-10]的采集范围，包括：(1)市级人口健康信息平台；(2)省级全民健康信息平台；(3)部(省)属医院信息平台。

针对市级人口健康信息平台的数据采集，主要包括业务运营数据，合理用药预警及管理数据，基础字典数据，临床诊疗记录相关数据等，见表 1。

针对省级全民健康信息平台的数据采集，主要包括业务运营数据、临床诊疗记录数据、基础字典数据、电子病历数据[11-12]，见表 2。

针对部(省)属医院信息平台的数据采集，主要包括医院信息系统(HIS)、实验室信息系统(LIS)、超声信息系统、病历信息系统等数据，见表 3。

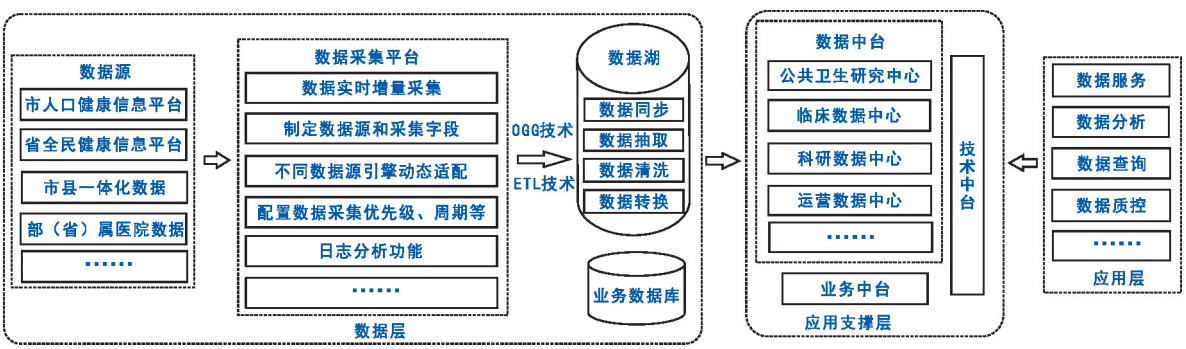

图 1　省级健康信息平台的数据采集流程

表1　市级人口健康信息平台数据采集内容示例

采集数据列	采集数据类明细
业务运营数据	医疗业务量(门诊量、床位数等)
	医疗收入
合理用药预警及管理数据	药品生产信息
	药品配送信息
	使用机构信息
	患者用药信息
基础字典数据	医院科室字典
	医护人员字典
	明细项目字典
	个人档案及家庭档案
	疾病管理数据(高血压等)
临床诊疗记录相关数据	服务对象信息
	就诊事件信息
	诊疗报告

表2　省级全民健康信息平台数据采集内容示例

采集数据列	采集数据类明细
业务运营数据	门诊挂号信息
	门诊收费信息
	住院结算信息
临床诊疗记录数据	门诊就诊信息
	住院就诊信息
	检查检验信息
	诊断明细
基础字典数据	项目/药品字典表
	科室字典表
	医护人员信息表
	医疗机构信息表
	医用设备表
电子病历数据	一般处置记录信息
	操作护理记录
	病程记录

3.2　数据采集规则

数据采集规则主要有3种：被动式数据采集解析、主动式数据采集解析和集中式数据采集解析[15-16]。被动式数据采集解析多用于定时数据采集[17]任务调用等场景；主动式数据采集解析多用于使用较为频繁的数据上报要求；集中式数据采集解析多用于基于平台的批量汇总数据上报等。

3.3　数据采集技术要求

数据采集平台需通过数据库同步技术对数据进行采集，应满足以下要求：(1)支持数据的实时增量采集；(2)支持指定数据源和采集字段；(3)可提供不同数据源引擎动态适配功能；(4)能提供数据优先级、采集时间、采集周期配置功能；(5)具有日志分析功能，可针对异常情况进行预警。

4　数据治理方案

4.1　数据清洗

将采集到的业务数据进行清洗，即对因不明原因导致的不规范、错误的字段信息自动进行统一的清洗，避免因部分明显错误导致上层应用服务的结论错误。

表3　部(省)属医院信息平台数据采集内容示例

采集数据列	采集数据类明细	采集数据类明细
医疗信息系统数据	医院信息系统(HIS)	放射信息系统
	实验室信息系统(LIS)	核医学信息系统
	超声信息系统	放疗信息系统
	病历信息系统	心电信息系统
	内镜信息系统	电子病历系统(EMR)
	移动护理系统	病案管理系统
	重症监护系统	手术麻醉系统
	财务系统	人事系统
	物流供应系统	资产系统

4.2　数据标准化和结构化处理

数据标准化处理遵循国内标准、国际标准、医疗行业标准、国际疾病指南标准等，包括ICD－10、ICD－9、HL7、CDA、医学主题词表(MeSH)、观测指标标识符逻辑命名与编码系统(LOINC)、药品词典规范－CFDA、ATC分类、医疗机构诊疗科目名录等。运用这些标准可对采集的数据自动进行标准化、归一化处理。数据结构化处理通过自然语义处理(NLP)技术，结合医疗专业术语的语义结构，将医疗语义信息自动按通用规则从自然语言表达扩展分析为结构化的Key－value模式，为后续的应用、挖掘、机器学习提供基础数据支持。

4.3　元数据处理

建立主数据和元数据管理机制。元数据管理平台的核心能力是以统一的数据标准对多源、异构的数据进行处理，形成统一、标准的大数据视图，通过对平台相关业务系统提供元数据服务，实现元数据的同步或匹配，包括用户账号权限、医生资质、组织机构、诊疗单元、服务单元、患者信息、检查项目、收费项目、药品目录等信息，以及用来描述主数据的关系数据，如组织机构与服务单元、组织机构与人员、服务单元与人员、检验检查项目和收费项目、临床诊断和标准ICD等的关系，以提高数据质量。

创建元数据之后，需要将其发布给医疗机构，以统一和规范各业务系统的主数据和业务数据，保证主数据编码的一致性、准确性。

4.4　数据质控管理

对多层数据的处理，采用定量加定性综合校验方法，运用多维质量监控、问题预警功能，协助人工智能平

台，实现数据的完整性、一致性、准确性、唯一性、及时性等。构建“数据采集——生产——治理——质量提升”的多层级医疗数据质控闭环，对于因技术原因导致的数据质量问题在源头即可进行纠正修复，对于因数据模型设计不合理导致的数据质量问题进行及时修复。

4.5　数据脱敏与加密处理

对敏感数据进行数据脱敏和加密处理[18]，自动去除或隐藏个人信息中的敏感信息（如患者姓名、身份证号、电话、地址等）[19]。通过脱敏或加密规则，实现敏感隐私数据的可靠保护，同时保持其他数据的可识别性和可用性[20]。

5　讨论

基于省级健康信息平台的数据采集及治理研究，与之前已有相关平台数据收集方法的区别在于，响应了新冠肺炎疫情的特殊背景，数据采集更符合国家对于疫情上报等重大公共卫生事件的数据上报要求，具有较好的针对性、可拓展性，是省级平台与国家平台数据采集信息化发展的产物。其主要优点在于：(1)整合了目前现有的各类平台数据，合理利用了现有资源，最大程度地实现数据的可利用性、完整性、综合性；(2)系统规划了基于省级健康信息平台的数据采集及治理方案，为省级平台与国家平台对接提供了规范的数据标准；(3)数据采集和治理方案实用性好，可操作性强，数据采集结合真实平台和真实数据，切实可行。

但是，本研究也存在一定不足与局限。对接的数据采集主要应用于疫情防控及重大公共卫生事件，对公共卫生数据针对性强，但不一定适用于其他场景。同时，主要探讨基于湖北省的省级健康信息平台，与全国其他省（自治区、直辖市）具体业务数据和业务内容不尽相同，并不能完全适用于其他省份。

需要说明的是，数据采集及治理是省级健康信息平台与国家平台对接的第一步，需要因地制宜，制定符合实际业务场景的数据采集及治理方案。只有前端的数据采集工作做到位，才不会影响后续业务层的数据管理决策需求。

同时，随着医疗健康大数据的发展，国家对数据采集的要求和管控会越来越精细化，与此同时数据采集的规范也会随之变化，在数据采集及治理方面如何最大可能地适应国家对医疗健康数据的管控要求，是未来需要持续关注的课题。

参考文献

[1]　陈　敏，肖树发，肖兴政，等.湖北省人口健康信息标准体系及基础代码规范研究[C]//湖北省人口健康信息技术交流大会论文集，2015.

[2]　中国医院协会病案管理专业委员会疾病分类分组.卫生部疾病分类与代码库形成[Z].2012.

[3]　国家卫生和计划生育委员会医政医管局.中国医疗服务操作项目分类及编码（征求意见稿）[Z].2014.

[4]　任宇飞，庹兵兵，杨　冲，等.线上线下一体化互联网医院云平台建设实践[J].中华医院管理杂志，2020，36（10）：837－840.

[5]　任宇飞，张晓祥，庹兵兵，等.集团化医院一体化管理与协同信息平台总体设计[J].中华医院管理杂志，2018，34(11)：932－935.

[6]　中华人民共和国卫生部.卫生信息化建设指导意见与发展规划(2011—2015)[Z].2010.

[7]　中华人民共和国国家卫生健康委员会.卫生信息数据元标准化规则：WS/T 303－2009[S].2009.

[8]　中华人民共和国国家卫生健康委员会.卫生信息数据模式描述指南：WS/T 304－2009[S].2009.

[9]　中华人民共和国国家卫生健康委员会.卫生信息数据元数据规范：WS/T 305－2009[S].2009.

[10]　中华人民共和国国家卫生健康委员会.卫生信息数据集分类与编码规则：WS/T 306－2009[S].2009.

[11]　中华人民共和国国家卫生健康委员会.卫生信息基本数据集编制规范：WS370－2012[S].2012.

[12]　中华人民共和国国家卫生健康委员会.卫生信息共享文档编制规范：WS/T 482－2016[S].2016.

[13]　中华人民共和国国家卫生健康委员会.卫生信息数据元目录：WS 363－2011[S].2011.

[14]　中华人民共和国国家卫生健康委员会.卫生信息数据元值域代码：WS 364－2011[S].2011.

[15]　中华人民共和国国家卫生健康委员会. 全国公共卫生信息化建设标准与规范（试行）：2020－12[S].2020.

[16]　中华人民共和国国家卫生健康委员会. 全国医院信息化建设标准与规范（试行）：2018－4[S].2018.

[17]　全国信息安全标准化技术委员会.信息安全技术网络安全等级保护基本要求：GB/T 22239－2019[S].2019.

[18]　全国信息安全标准化技术委员会.信息安全技术 网络安全等级保护安全设计技术要求：GB/T 25070－2019[S].2019.

[19]　全国信息安全标准化技术委员会.信息安全技术网络安全等级保护测评要求：GB/T 28448－2019[S].2019.

[20]　全国信息安全标准化技术委员会.信息安全等级保护管理办法：公通字〔2007〕43号[S].2007.

通信作者：
汪火明：华中科技大学同济医学院附属同济医院计算机中心副主任
E－mail：9112127@qq.com

收稿日期：2021－06－20
修回日期：2021－10－23
责任编辑：姚　涛