

面向医院的大数据治理模型设计

俞鹏飞, 罗颖文, 刘建模, 易应萍

(南昌大学第二附属医院信息处医疗大数据研究中心, 江西 南昌 330000)

摘要:目前,促进健康医疗大数据发展已经成为我国健康产业发展的重要国策,但是国内医院普遍存在医疗数据使用困难、数据质量不高等问题。为此,本文提出了一种基于大数据架构的医院数据平台建设方法及治理模型。阐述了当前医疗数据面临的问题及其原因,总结数据仓库建模理论与医疗数据特点,分析医疗数据治理模型框架,旨在通过分层分域组织医疗数据,提升数据质量、加速数据应用,实现医院医疗数据资产化运营,促进大数据平台为医院临床、管理、科研发展赋能。

关键词:医疗大数据;数据模型;医院数据平台

中图分类号:R197

文献标识码:B

DOI:10.3969/j.issn.1006-1959.2021.10.005

文章编号:1006-1959(2021)10-0018-03

Design of a Hospital-oriented Big Data Governance Model

YU Peng-fei, LUO Hao-wen, LIU Jian-mo, YI Ying-ping

(Medical Big Data Research Center, Information Division, the Second Affiliated Hospital of Nanchang University, Nanchang 330000, Jiangxi, China)

Abstract: At present, promoting the development of health and medical big data has become an important national policy for the development of my country's health industry, but domestic hospitals generally have problems such as difficulty in using medical data and low data quality. To this end, this paper proposes a hospital data platform construction method and governance model based on big data architecture. It expounds the current medical data facing problems and their causes, summarizes the data warehouse modeling theory and medical data characteristics, analyzes the medical data governance model framework, the aim is to organize medical data in layers and domains, improve data quality, accelerate data application, realize the asset operation of hospital medical data, and promote the big data platform to empower the hospital's clinical, management, and scientific research development.

Key words: Medical big data; Data model; Hospital data platform

2016年6月,国务院办公厅发布《关于促进和规范健康医疗大数据应用发展的指导意见》^[1]。同年底,国家卫生计生委启动健康医疗大数据中心与产业园建设国家试点工程,全面推动大数据应用与健康医疗行业的深度融合。目前健康医疗大数据已成为国家大数据战略重要组成部分。同时,医院是数据密集产生的源头,且医疗数据、健康数据具有非常高的应用价值,通过大数据技术能够挖掘出重大价值,为临床诊疗、管理运营、医疗科研赋能。但由于数据质量低、标准化难等问题,导致医院数据使用效率低、难以产生价值。因此,建立医院大数据平台以及设计数据治理模型,利用数据仓库建模理论与医院数据平台建设经验^[2],起到提升数据质量、提高开发使用效率的作用,将使得医院数据资产化、促进智慧医院建设发展^[3]。为此,本文提出了一种基于大数据架构的医院数据平台建设方法及治理模型,分析如下。

1 医疗数据的现状与挑战

1.1 数据种类多 不同于传统临床数据中心仅存储临床系统产生的原始数据,大数据平台采集的数据

基金项目:1.国家自然科学基金项目(编号:81960609);2.国家重点研发计划(编号:2018YFC1312902);3.江西省重点研发计划(编号:2018ACH80004)

作者简介:俞鹏飞(1992.1-),男,江西南昌人,硕士,初级工程师,主要从事医疗大数据与人工智能研究

通讯作者:易应萍(1963.4-),女,江西新余人,本科,研究员,主要从事医疗大数据与人工智能、公共卫生管理研究

类型、种类更多,需要将不同系统数据进行集成汇聚^[4]。主要包含医嘱、药品、检验、手术治疗等结构化数据,以及病历、护理文书、检查病理报告等非结构化、半结构化数据、基因测序数据、医学影像文件数据等,且需要通过自然语言处理对非结构化数据进行信息提取、结构化处理。

1.2 数据质量不高 医疗数据质量普遍不高,主要体现在完整性、规范性、整合性。一方面,医院数据采集自各个业务系统,为了保障业务运行效率,采用前端验证后录入方式较难,无法保证数据完整性。另一方面,医院诊断、用药等医学术语标准多、更新快,不同医护人员录入习惯也不一致,导致数据未按统一标准录入。同时,医院数据质控体系不完善,除科研需求外,医生很难有动力完善病历记录,也缺乏相关意识。建立质控点必然需要改造业务流程,使得系统操作更繁琐,很难在临床系统实施。

1.3 数据开发难 大部分医院没有统一的数据开发平台,数据均以原始状态存储在各个系统数据库中。开发一项数据应用时,需要提取多个接口数据,即使医院已经有服务总线,数据提取、处理也要花费大量时间^[5]。各应用数据开发过程相互独立,抽取、清洗、处理过程需要重复开发,结果无法重用,导致数据应用效率低,成本高。尤其是基于临床诊疗数据的应用,难以形成从数据采集、存储、整合、分析到应用的完整闭环。应用过程中,难免会出现数据采集不完整、数量质量不高、信息提取准确等问题,极大降低

了开发效率。

2 医疗大数据治理模型与应用

目前医疗数据存在的问题严重制约了医院在智慧化建设过程中必须的数据开发能力,为了解决数据汇聚难、治理不足、开发效率低等问题,本文提出基于大数据平台的医疗数据分层分域治理模型。首先,在汇聚层搭建数据集成框架、建立严格的数据质量控制系统,及时发现、解决数据质量问题。然后建立数据分层分域模型,对医疗数据进行归纳整理,形成科研域、管理域、临床域数据,并建立了包括汇聚层、主题层、汇总层、应用层的数据模型,解决大数据平台数据量大、数据间关系复杂、数据不一致等问题。

2.1 数据集成框架 大数据平台数据汇聚包括对不同类型、不同来源、不同时间的数据接入。对于结构化数据,按照数据汇聚的传输方式,可以分为文件传输、数据抽取、消息推送等方式。其中文件传输方式需要业务系统定时进行数据抽取,需进行业务系统改造;数据抽取不需要业务系统改造,适用场景多,需要解决的关键问题有多数据源适配、增量数据抽取、数据一致性审查等。我院数据汇聚采用开源 ETL 工具 Kettle 实现多数据源适配,通过配置不同数据库连接,实现对不同数据库进行数据抽取任务的创建、运行、运维,有效提高了数据汇聚效率,减少运维成本^[6]。通过 ETL 平台抽取的数据需要在抽取过程中完成数据脱敏、加密存储以及一致性校验。按照 HIPAA 中定义的关键隐私数据(姓名、身份证、联系方式、家庭住址、生物信息等)通过加密算法计算后导入大数据平台。数据抽取流程框架见图 1。

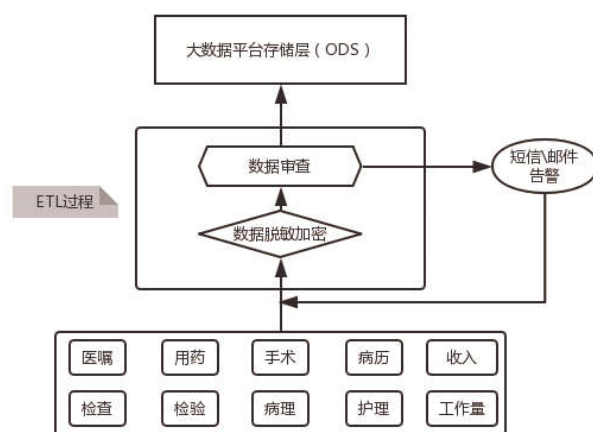


图 1 数据抽取流程框架

为了确保数据抽取的一致性,需要建立数据审查机制。我院分别对于历史数据、实时数据设计不同抽取流程,并且通过记录日志、实时警告等方式保证数据完整、正确地将接入平台。对于实时数据采用运行较快的方法,在保障数据不缺少的基础上,验证数据一致性。对于历史数据按照不同数据类型,选取逻辑检查方法定期生成数据审查报告,确保数据一致。数据审查机制内容见表 1。

2.2 数据治理框架 除了院内产生的业务数据,医疗数据还包括患者的体检数据、医保数据、随访数据、家庭健康监测数据等。为了对这些数据进行梳理,使不同模块间耦合度降低,提高利用效率,我院建立了分层分域数据治理模型,见图 2。该模型将数据分为临床域、科研域、管理域,由下而上建立数据源接口层、数据主题层、数据汇总层和数据应用层。

表 1 大数据平台数据汇聚审查方法

检查方法	检查描述	使用案例
数量检查	对源数据与抽取后数据进行数据量检查	通用
值域评判	对数值、枚举类型数据进行取值范围检查	诊断、医嘱等基础信息
关联性检查	定义不同数据相关性,检查是否满足相关性特征	工作量与收入关联等
平衡性检查	通过对不同数据进行简单运算,检验数据是否平衡	门诊住院收入与总收入

数据源接口层负责组织管理多源数据汇聚,即数据的采集、转换、存储,采用分布式文件系统存储保存加密、脱敏后的基础数据。通过数据审查方法保证数据一致性、唯一性、正确性等要求,以尽量少的代价检测与源数据的一致性。

数据主题层将接口层存储的数据经过统一清洗、编码转换、整合后形成主题域。其主要的功能是设计好主题域下模型划分。该层次的数据模型的目标是灵活地表达业务过程,将源系统关系型的数据结构,按照主题划分整合,将大概率一起使用的数据整合到统一主题域中。如源系统中医嘱信息通常包括医嘱项、医嘱记录、医嘱执行记录等数据表用于记录医嘱的不同数据信息,而在主题层则将医嘱相关

数据进行主题化处理,提取事实表与维度表,建立医嘱主题等。

数据汇总层及数据应用层则面向应用进行数据处理,对相关业务来说,每次处理明细数据速度慢、代价高,在汇总层将明细数据进行有效汇总,提供临时数据挖掘使用,同时加快应用层调用时的速度。

在应用层则形成标签集、指标集、应用宽表提供外部数据共享。

医疗标签集由患者画像特征化标签、统计类标签、预测分析标签组成,如患者基本信息、平均费用、就诊频次、疾病诊断路径等标签,是能够描述患者健康信息的集合^[7]。通过对患者进行标签化特征描述,能够方便临床科研分析以及建立人工智能预测模型。

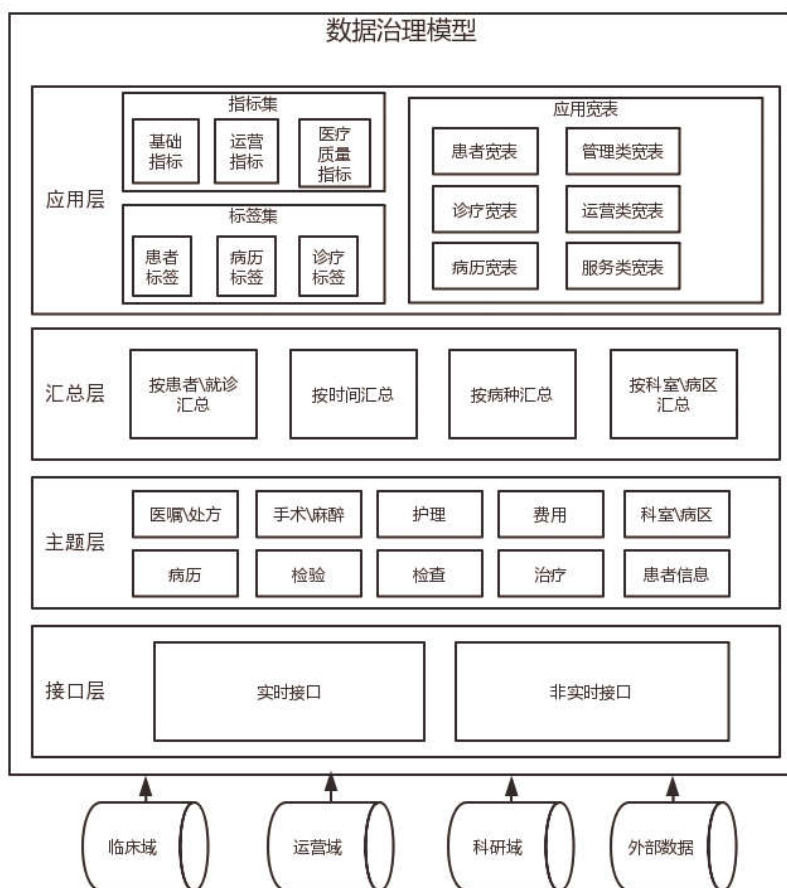


图2 数据治理模型

指标集则是面向医院管理运营的数据服务,通过将医院运营过程中各类统计数据实时产生相关指标,汇聚之后能够全面反应医院运营情况,如门诊人次、住院人数、平均住院天数、传染病诊断等。应用宽表是面向外部应用而建立多字段数据表,主要用于进行数据访问控制、降低数据复杂度、减少数据交互及加速数据应用的作用。

3 应用实践

通过建立医院大数据平台以及数据治理模型,江西省某三甲医院已经从院内历史使用及正在使用的 74 个医疗业务系统中汇聚了 2006 年~2020 年的所有数据,结构化数据总数据量达到 11.4 亿条,非结构化数据完成部分病例、检验检查报告文本结构化处理。经过数据清洗、结构化、标准化处理后,形成 9.1 亿条标准化数据。通过建立数据分层分域治理模型,建立了 15 个主题域、数百患者标签以及运营指标,支撑了医院临床科研大数据平台、运营管理 BI 系统、临床辅助决策系统等大数据应用。支撑医院科研人员快速检索历史数据,医院管理者实时直观了解医院运营状况,为临床工作者提供智能化辅助诊疗。

4 总结

医疗大数据已经成为国家重要发展战略,充分挖掘利用医院数据对医学科研发展、提高医院运营

管理效率、提高医疗质量都有重大意义。搭建医院大数据平台,利用数据治理模型对数据进行汇聚、处理,能够提升医院的数据应用能力,发挥数据价值。但目前医院大数据平台发展时间较短,相关研究与应用仍不成熟,应在建设过程中不断探索、升级,实现医院数据资产化、智能化。

参考文献:

- [1]国务院.关于促进和规范健康医疗大数据应用发展的指导意见[EB/OL].[2019-08-05].http://www.gov.cn/zhengce/content/2016-06/24/content_5085091.htm.
- [2]刁琰.基于临床数据中心的医院信息集成平台建设[J].医学信息学,2018,31(24):15-18.
- [3]熊旭峰.浅谈智慧医院建设[J].信息周刊,2018(8):331.
- [4]马少锋,温锋,陈超.基于数据仓库的医院数据分析平台建设与应用[J].医学信息学杂志,2017,38(7):18-21.
- [5]黄跃.患者全息视图系统的构建与应用[J].中国数字医学,2019,14(2):79-81.
- [6]刘蕾,廖茂成,李韶朗,等.基于 Caché 数据库和 ETL 过程的医疗质量辅助决策方法研究[J].中国卫生质量管理,2015,22(1):94-96.
- [7]周晓英.电子健康档案:特征、构成和标准化问题研究[J].中国国情国力,2017,6(2):85-90.

收稿日期:2020-12-21;修回日期:2021-01-04

编辑/钱洪飞