·智慧医疗· doi:10.3969/j.issn.1671-8348.2022.24.029
网络首发 http://kns.cnki.net/kcms/detail/50.1097.R.20221014.1454.008.html(2022-10-14)

医院数据存在问题与管理对策研究*

周 琳,王 飞,赵浩宇△
(陆军军医大学第一附属医院医学大数据与人工智能中心,重庆 400038)

[摘要] **目的** 根据该院大数据平台实践经历,探讨医疗数据管理的难点与对策。**方法** 通过分析该院数据存在的问题,遇到的困难,探讨构建大数据平台对医院数据进行有效汇聚、治理及管理,完善数据管理对策。**结果** 该医疗数据具有数据规模大、数据类型繁多、数据流转速度极快、价值密度较低四大特性。通过平台建设、数据整合、数据处理、数据质量管理完成对医院数据的汇聚并提出管理对策。**结论** 医院构建大数据平台能完善医疗数据管理,可为医院科研的发展、医疗模式转型和医疗领域创新应用带来机遇和动力。

[关键词] 医疗大数据;数据管理;对策
[中图法分类号] R197.32 [文献标识码] A [文章编号] 1671-8348(2022)24-4300-04

Existing problems in hospital data and management countermeasures*

ZHOU Lin, *WANG Fei*, *ZHAO Haoyu*△
(*Medical Big Data and Artificial Intelligence Center*, *First Affiliated Hospital of Army Military Medical University*, *Chongqing* 400038, *China*)

[**Abstract**] **Objective** To explore the difficulties and countermeasures of medical data management according to the practical experience of big data platform in this hospital. **Methods** By analyzing the existing problems and difficulties of data in this hospital, a big data platform was constructed to effectively gather, govern and manage the hospital data, and the data management countermeasures were perfected. **Results** The medical data in this hospital have the four characteristics of large data scale, various data type, very fast data flow speed and low value density. Through the platform construction, data integration, data processing and data quality management, the hospital data aggregation was completed and the management countermeasures were put forward. **Conclusion** The construction of hospital big data platform can perfect the medical data management, and bring the opportunities and impetus for the development of scientific research, transformation of medical model and the innovation and application of medical field.

[**Key words**] medical big data; data management; countermeasure

我国医疗数据资源丰富,尤其是大型三甲综合医院,有着几十年的医疗数据沉淀。随着医院信息化的发展,医院对数据的管理能力也有了一定的提高。但受困于数据架构的局限,医疗数据质量参差不齐,大量数据未被有效利用。传统数据库在数据整合、数据管理、数据治理、数据利用等方面的使用效率不高,面对临床、科研、管理、患者自身持续增长的业务需求得不到及时满足,难以用数据支撑临床科室的科研创新研究。如何建设一套新的平台架构,充分利用现有医疗数据,通过对数据的抽取、整合和治理,解决当前数据管理的问题,提高医疗领域的科研创新能力,满足医疗模式转型的应用需求成为近年来医院信息部门研究的一个重要方向[1]。

1 数据平台的建设与管理

1.1 医院数据存在的问题

本院医疗数据具有数据规模大、数据类型繁多、数据流转速度极快、价值密度较低四大特性。由于临床病历书写的质量参差不齐、各系统数据结构类型存在差异、数据整合难度大、数据库在数据处理(特别是复杂组合条件下的查询)方面执行效率太低、数据治理更是无从谈起。数据使用效率一直不高。如何提升数据质量及数据使用效率,为医院疾病诊断相关分组(diagnosis related groups,DRG)开展和临床科研服务成为本院当前信息化重点建设的内容之一[2]。

* 基金项目:国家重点研发计划项目(2018YFB2101204)。 作者简介:周琳(1974-),高级工程师,硕士,主要从事医学大数据研究。
△ 通信作者,E-mail:153591907@qq.com。

1.2 数据平台比较

本院现有数据平台采用的是传统关系数据库，其建立在关系模型基础上，借助于集合代数等数学概念和方法处理数据库中的数据，在实时性、一致性，以及对结构化数据处理等方面均具有自身优势。本院发展信息化较早，在关系数据库建设方面具有一定经验。近年来，随着大数据技术不断成熟，使用领域越来越广，利用大数据进行医学方面的研究也更加深入。基于 Hadoop 平台下的数据库采用 shared-nothing 架构，每个节点均有自己的操作系统、数据库和硬件资源，节点之间通过网络来通信。该平台能整合不同类型数据，并可以对数据进行集中清洗和治理，通过 HBase 组件支持对实时数据的读写处理，通过 Hive 组件构建数据仓库，舍弃了索引、关系以及事务处理等关系型数据库的特性，在数据查询和处理方面效率均有了极大的提升[3]。通过比较，本院决定从创新技术入手，选择建设基于 Hadoop 架构的大数据平台下的数据库提升医院数据质量和处理能力。关系数据库和大数据平台数据库二者之间的主要区别，见表 1。

表 1 RDBMS 和 Hadoop 技术对比分析

项目	RDBMS	Hadoop 平台
容量规模	GB→TB	TB→PB 以上
数据模型	结构化数据	结构化、半结构化、非结构化
架构方式	存储集中，计算集中，纵向扩展	分布式存储，分布式计算，横向扩展
查询语言	SQL	HQL
数据存储位置	Raw Device 或者 Local FS	HDFS
处理性能	数据量达到 TB 级别便达到性能瓶颈	分布式处理，性能较传统方式提升几倍至几十倍
执行	Executor	MapReduce
应用价值	侧重数据的操作性，兼顾统计报表	关注数据的业务决策价值，强调数据挖掘与综合分析
可靠性	取决于关键节点，大数据量的备份和恢复较为困难	海量数据的 3 副本冗余备份，关键组件都提供 HA 功能
经济性	服务器配置要求高，软件采购成本及后续服务费高	运行在普通 x86 服务器上，软件成本低
扩展性	弱，扩展时需要进行大规模的数据迁移，成本高	强，只需增加节点而不需要影响原集群中的数据
一致性	强	一般
实时性	强	一般
更新	多次读/写	一次写入，多次读取

1.3 大数据平台建设

大数据时代医院对数据的利用已从简单的报表分析走向可预测分析阶段，及时、准确的数据是进行数据加工和分析的基础，想要利用好医院的数据进行基于大数据平台的系统开发，为科学预测和政策制定提供技术支持[4]，数据的有效同步是平台实施的关键步骤。本院主要采用 ORACLE 和 SQL Server 数据库自带的数据同步功能来完成平台数据的实时同步。对于 ORACLE 数据库，采用 OGG 方式进行增量数据的同步。其中 Extract 进程运行在源系统上，负责捕获数据更改。Replicat 运行在目标计算机上，负责将更改应用于目标数据库。而源系统和目标系统之间则主要以 Trail 和 Flat 文件来进行数据传输[4]。对于 SQL SERVER 数据库，采用 CDC 方式，基于日志抓取，识别出变化的数据，获取增量数据。架构中前置机部署在医院内网，通过抽取工具 sqoop 抽取增量数据到 hive 库，再通过脚本合并增量数据为全量。业务库到前置库之间，通过捕获日志进程抓取变更日志，并将变更日志同步到前置库，通过前置库的解析日志功能，解析出数据的变更。平台整体架构见图 1。

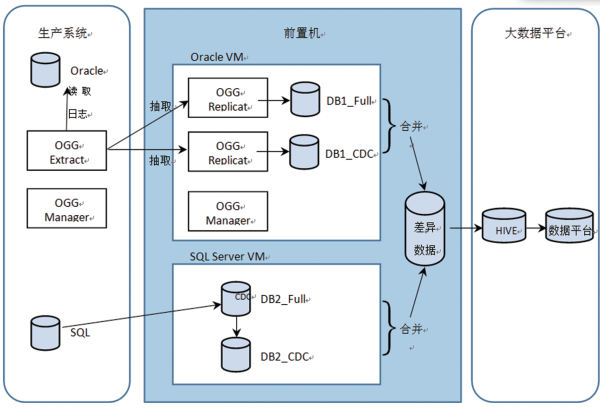

图 1 大数据平台架构示意图

1.4 管理对策

本院经过 20 多年的数字化医院建设积累了海量的医疗数据，但医院信息系统（HIS）、实验室信息系统（LIS）、影像归档和通信系统（PACS）、放射科信息系统（RIS）、电子病历系统（EMRS）、移动医护、急诊、心电、手术麻醉、体检等各系统数据分散存放，结构互

异，整合难度较大。通过基于 Hadoop 的大数据平台建设对当前数据进行有效整合和管理，解决当前数据存在的问题。

1.4.1 异构数据整合

（1）医院的医疗系统中沉淀着大量基础数据，这些数据产生于不同历史时期，来源也不同，标准也不统一，本院也经历过新老药品切换、物价调整、疾病编码版本升级等基础数据变动的情况；（2）医院内系统众多，各系统数据库数据类型各异，有 DB 结构化数据、XML 文档数据、DICOM 影像数据、动态影像数据、自由文本文档数据、PDF 文档数据等；（3）数据来源多种多样，有 HIS 基础诊疗数据、EMRS 电子病历数据、PACS 放射类检查数据、LIS 检验数据、康复管理数据、重症监护数据、病理影像数据等；（4）数据存储方式多样，有在线数据、近线数据、历史归档数据等。如此多的“异源异构”数据给数据的汇聚与利用带来很大难度[5]。大数据平台能通过抽取、转换、装载（ETL）工具，利用数据库 OGG 和 CDC 功能将异构异源的数据抽取到大数据平台，将数据文件切分成若干个块，存储到不同节点上，最后以 textfile 格式进行存储。通过这一步的数据抽取、整合和处理，将医院所有数据进行了有效汇聚。如信息部门以前为临床科室查询数据时需要在多个系统及数据库里查询数据，现在只需要在一个平台输入条件即可查询，避免了以前多的数据库查询带来的麻烦，也不会占用医院在用生产库资源，提高了数据使用效率和数据使用安全[6]。

1.4.2 数据结构化处理

医院除数据库存储的是结构化数据外，其他很多均为非结构化数据，类型五花八门，但只有转换为结构化数据才便于分析，而且结构化数据字段中依然存在大量需要进一步语义识别的自由文本数据；在 EMRS 中医师的电子病历文书记录蕴含了大量的有价值信息，一般都是以自由文本或半结构化数据存储；影像等系统则存储的是 DICOM 格式的数据。利用平台处理非结构化数据首先要将数据进行抽取汇聚，再利用平台技术对数据进行文本的词汇切分、词性分析、歧义处理等实体提取，然后对词汇相关度、句子相关度、篇章相关度、句法分析进行语义处理，最后建立向量空间模型、主题模型[7]。

1.4.3 数据质量提升

数据质量问题是目前数据处理过程中遇到的普遍现象。各分系统存在数据标准不一、数据内涵错误等问题；EMRS 中医师病历书写未按照 ICD10 标准，每名医师均有自己的书写习惯，如诊断名称不统一、相同疾病描述有差异等；口话太多导致的垃圾数据；部分字段数据缺失不完整等都是造成医院数据质量不高的原因。提升数据质量除规范医疗质量管理外，还必须借助大数据平台对数据进行清洗和治理，在此基础上进行精细化、细粒度数据分析，如结构化处理、字段切词、归一化处理等，以此提升医疗数据质量[8]。如当在数据库搜索诊断名称为“脑梗死”的疾病，但不同医师在病历书写时会将这种诊断名称写成如“脑梗死、出血性脑梗死、脑干梗死、大脑动脉栓塞引起的脑梗死、大脑动脉血栓形成引起的脑梗死”等，可能都是描述的这一疾病。如在以前只能在数据库里通过模糊查询一个一个搜索，借助现有平台只需要输入一个“脑梗死”，系统会将以上进行了结构化和归一化处理的相关诊断名称默认为是同一个诊断名称，并返回所有数据，这样既方便了信息部门查询数据，也保障了临床科研数据质量的准确。

1.4.4 数据分析效率提升

本院传统的关系数据库面对超大表、多表数据分析时存在较大性能瓶颈，执行效率低，返回结果慢，而且在执行数据查询时还会占用医院在用生产库资源，影响医院业务系统运行。特别是随着医院在大数据及人工智能领域的摄入研究，以后深度学习等人工智能算法对图形处理器（GPU）处理要求越来越高，当前数据分析已经难以满足医院使用需求[9]。新建大数据平台采用分布式存储和分布式计算，通过内部的资源管理和调度系统合理分配资源，结合 MapReduce 分布式离线计算框架，可以对数据进行多进程同步运算，提升数据分析效率。如以前信息部门在为临床科室提供科研数据服务时，当面对多条件的复杂组合查询时由于涉及的业务数据表太多，执行效率很低，不仅会占用数据库资源，返回数据结果也很慢，如果再遇到时间区间跨度达 5 年以上的、历史库和在用生产库数据没有整合更加难以完成。借助大数据平台只需要通过添加纳排条件，以分布式计算为基础，即可实现秒级查询和数据展现。

2 应用效果

2.1 保障了数据完整性

高质量数据来源于数据收集，是数据设计及数据分析、评估、修正等环节强有力的保证[5]。通过抽取数据将不同类型不同来源的数据进行了有效整合，不重复也不会遗漏，最大限度地保持整个医疗数据的完整性。

2.2 促进了数据标准化

医疗过程中存在着大量专业术语和专业定义，但因为种种原因，医院医疗系统及医疗工作者书写的病历中却存在大量的非标准化数据，数据表达方式随意性较大。通过平台对数据的清洗和治理，最大限度地保障了医疗数据的规范性和标准化。

2.3 提升了数据质量

平台能够通过数据清洗，结构化、归一化处理，词汇切分，语义处理等，将原来非标准的、不完整的数据进行规范化，提升了医院医疗数据质量，完善了医院医疗质量管理。

2.4 催生智能化应用落地

通过大数据平台处理和汇聚的医疗数据并非只是为了科研、教学或管理等场景使用，更多地是为后续医院在智能化方面的建设打下基础。首先需要把这些散落的数据整合成为标准的患者诊疗模型，完成诊疗模型构建和数据处理，再根据这些整合数据，通过人工智能学习，构建智能辅诊系统，预测出患者下一步的健康变化，自动推荐诊断和治疗方案，实现更大的医疗数据生态，催生更多医学领域的智能化应用落地[10]。

综上所述，我国许多大型医院经过几十年信息化发展，积累了大量的临床诊疗数据，这些数据以前在管理和使用上还不够规范和完善，随着大数据技术的不断发展成熟，医院构建大数据平台能够将这些数据进行有效汇聚和管理，提升医院诊疗数据质量，为医院临床科研的支撑、诊疗模式的转型和医疗领域的智能化应用带来机遇和动力。由于医院在大数据和人工智能领域的建设起步较晚，相关技术人员较少，经验不足，在大数据建设和管理上还会遇到很多新的难点和痛点。只有通过对新技术新业务的不断学习，掌握更为有效的数据采集、清洗、加工工具和方法，在实践中不断摸索和发展，才能建设好医疗科研大数据平台，为医院创新发展打下坚实的基础[11]。

参考文献

[1] 郑西川，陈霆，傅一旻，等. 医疗机构大数据分析功能及应用策略研究[J]. 中国数字医学，2018，13(3)：13-15.

[2] 倪晓华. 利用 GATE 的 XML 配置文件实现病历短语抽取的机器学习方法[J]. 中国医疗设备，2017，32(7)：124-125.

[3] 赵广智. Hadoop 与关系型数据库在电信行业中的应用研究[J]. 广东通信技术，2017，37(9)：40-46.

[4] 刘晓亮，王坤，马军，等. 大数据时代的卫生信息化建设思考[J]. 中国卫生信息管理杂志，2014，11(1)：43-46.

[5] 米春香. 大数据技术助力医院精细化管理[J]. 中国医疗设备，2019，34(7)：93-95.

[6] 王才有. 大数据时代的医院数据平台建设[J]. 中国医院，2016，20(1)：15-17.

[7] 刘金晶，曹文洁. 大数据环境下的数据质量管理策略[J]. 软件导刊，2017，16(3)：176-178.

[8] 王笑笑，赵飞，梁志金，等. 基于数据挖掘技术的护理学研究现状[J]. 解放军护理杂志，2019，36(8)：59-62.

[9] 王韶锋，赵善斌，杨静，等. 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.

[10] 汪鹏，吴昊，罗阳，等. 医疗大数据应用需求分析与平台建设构想[J]. 中国医院管理，2015，35(6)：40-42

[11] 王红迁，汪鹏，王飞，等. 保障医疗大数据安全及其实践[J]. 医学信息学杂志，2017，38(12)：43-47.

（收稿日期：2022-02-23 修回日期：2022-06-28）

（上接第 4299 页）

[19] HLLEIN A，ZHAO M，SCHABATH R，et al. An artificial intelligence (AI) approach for automated flow cytometric diagnosis of B-Cell lymphoma[J]. Blood，2018，132：2856.

[20] KO B S，WANG Y F，LI J L，et al. Clinically validated machine learning algorithm for detecting residual diseases with multicolor flow cytometry analysis in acute myeloid leukemia and myelodysplastic syndrome[J]. E Bio Med，2018，37(1)：91-100.

[21] HOLZINGER A，LANGS G，DENK H，et al. Causability and explainability of artificial intelligence in medicine[J]. Wiley Interdiscip Rev Data Min Knowl Discov，2019，9(4)：e1312.

[22] AMANN J，BLASIMME A，VAYENA E，et al. Explainability for artificial intelligence in healthcare：a multidisciplinary perspective[J]. BMC Med Inform Decis Mak，2020，20(1)：310.

[23] CUTILLO C M，SHARMA K R，FOSCHINI L，et al. Machine intelligence in healthcare-perspectives on trustworthiness，explainability，usability，and transparency[J]. NPJ Digit Med，2020，3(1)：47.

（收稿日期：2022-04-21 修回日期：2022-09-11）