大数据背景下医院数据质量评价模型及监控管理模式探索

冯晨阳 刘迷迷 刘 强

（中山大学肿瘤防治中心 广州 510006） （中山大学中山医学院 广州 510080） （医渡云（北京）技术有限公司 广州 510006）

何仲廉 周 毅 李超峰

（中山大学肿瘤防治中心 广州 510006） （中山大学中山医学院 广州 510080） （中山大学肿瘤防治中心 广州 510006）

〔摘要〕 结合中山大学肿瘤防治中心基于医疗大数据平台进行数据质量评价及监管的实践经验，阐述该模式应用的技术框架，包括大数据平台和数据质量评价模型构建方法以及针对各维度量化评分计算方法，分析该模式应用的阶段性成果、成效与创新情况，为医院数据质量评价提供可行的模式和思路。

〔关键词〕 数据质量评价；数据治理；医疗大数据平台

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.07.005

Exploration of Hospital Data Quality Evaluation Model and Monitoring Management Mode under the Background of Big Data

FENG Chenyang, Sun Yat－sen Cancer Center, Guangzhou 510006, China; LIU Mimi, Zhongshan School of Medicine, Sun Yat－sen University, Guangzhou 510080, China; LIU Qiang, Yidu Cloud (Beijing) Technology Co. Ltd., Guangzhou 510006, China; HE Zhonglian, Sun Yat－sen Cancer Center, Guangzhou 510006, China; ZHOU Yi, Zhongshan School of Medicine, Sun Yat－sen University, Guangzhou 510080, China; LI Chaofeng, Sun Yat－sen Cancer Center, Guangzhou 510006, China

〔**Abstract**〕 Combined with the practice experience of Sun Yat－sen Cancer Center in data quality evaluation and supervision based on the medical big data platform, the paper elaborates the technical framework of the application of this mode, including the construction method of the big data platform and the data quality evaluation model, and the calculation method of quantitative score for each dimension, analyzes the phased results, effectiveness and innovation of the application of this mode, and provides feasible mode and thought for hospital data quality evaluation.

〔**Keywords**〕 data quality evaluation; data governance; medical big data platform

1 引言

随着大数据、人工智能等新兴技术的发展，医

〔修回日期〕 2022－06－22

〔作者简介〕 冯晨阳，工程师，发表论文 3 篇；通信作者：李超峰，高级工程师。

疗行业数据量呈爆炸式增长。政府部门、医疗机构已逐渐意识到数据作为一种资产的重要性及其蕴含的巨大价值[1]。医疗大数据是指医疗行业中日积月累产生的数据[2]，具有类型繁杂、多源异构、非结构化数据占比高等特点，对数据利用带来挑战[3]。要开展数据治理工作应先了解所存在问题，对数据质量做全面性评价。目前针对医院数据质量评价的标准和体系研究已较为成熟，其中大多数研究均将数据完整性、一致性、规范性、准确性、时效性作为评价数据质量的重要维度[4-5]。但在医院信息化项目的实践过程中发现，基于医疗业务流程的复杂性、不同机构的信息系统差异性，不同医疗机构存在的数据问题和治理的侧重点均不同，如何将标准化的评价模型与院内实际情况相结合，从而因地制宜对院内数据质量进行客观真实评价是值得探讨的问题。本文将中山大学肿瘤防治中心（以下简称中肿）基于医疗大数据平台进行数据质量评价及监管的应用情况进行阐述，并提出相关思考和建议。

2 技术框架

2.1 大数据平台构建

大数据平台基于分布式技术架构，在不影响业务系统的前提下，将包括医院信息系统（Hospital Information System，HIS），电子病历（Electronic Medical Record，EMR），检验信息系统（Laboratory Information System，LIS），医学影像存储与传输系统（Picture Archiving and Communication System，PACS）等在内的多个系统数据整合到标准的数据模型，形成患者维度和就诊维度的全景数据。同时，大数据平台通过自然语言处理（Natural Language Processing，NLP）技术将医院内业务系统产生的原始医疗文本数据进行结构化处理。大数据平台整体框架，见图 1。

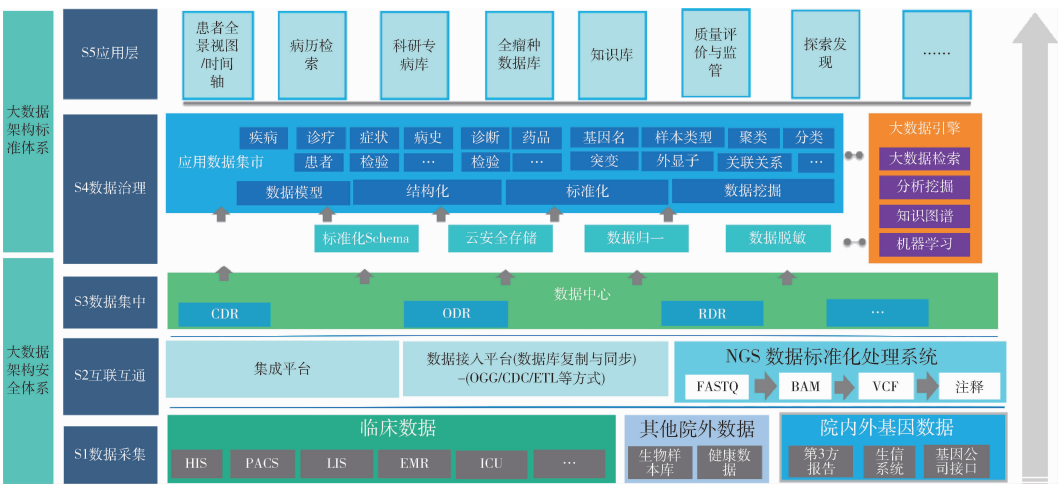

图 1 大数据平台整体框架

2.1.1 数据汇聚采集 在医院内网环境配置数据同步前置机，通过 OGG 和发布订阅等数据库同步技术，实现院内临床诊疗、医技等多源异构数据全量接入，支持以 T+0/T+1/T+7 等不同频度进行更新和调度。汇聚来自 HIS、EMR、LIS、PACS、护理、病理、超声、放疗、手麻、重症、内镜、心电图、生物样本库、体检、肺功能等系统的业务数据。同时前置机同步时会采用数据加密标准（Data Encryption Standard，DES）加密算法对患者姓名、身份证号、电话、地址等敏感信息进行脱敏处理，见图 2。

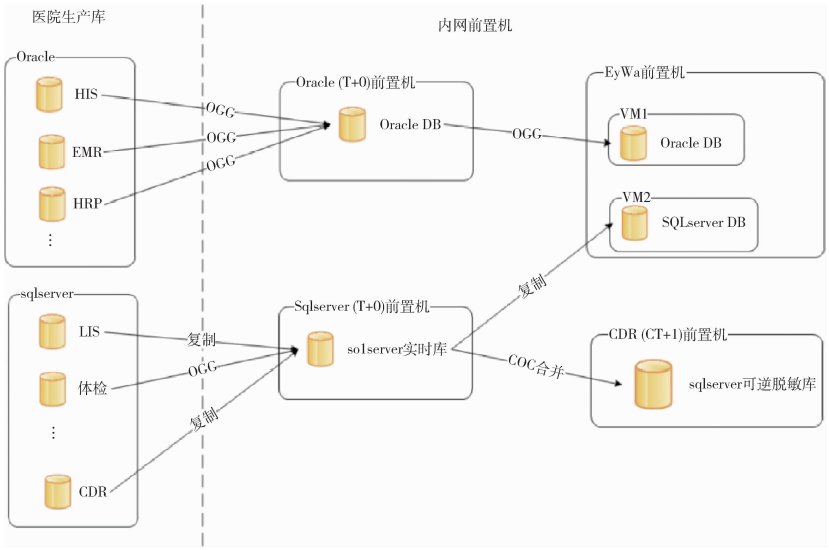

图 2 数据库同步

2.1.2 数据处理加工 （1）数据清洗。对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正错误并保证数据一致性。一致性检查是根据每个变量的合理取值范围和相互关系检查数据是否合乎要求，是否存在超出正常范围、逻辑上不合理或者相互矛盾的数据，如无效值和缺失值、残缺数据、错误数据、重复数据等。（2）患者主索引（Enterprise Master Patient Index，EMPI）合并和关联。通过 EMPI 整合患者历次就诊数据，建立患者维度数据集。根据患者身份证号、就诊卡号等可直接确认患者身份的数据字段做就诊关联并生成 EMPI 患者唯一标识。当无法直接确认患者身份时，根据姓名、性别、出生日期、身份证号等数据并通过 EMPI 算法规则判断是否为同一患者。通过 EMPI 可将患者单次检查或检验、就诊维度进行关联。（3）标准化归一。基于相关指南和行业等标准，如药品词典规范解剖学、治疗学及化学分类系统（Anatomical Therapeutic Chemical，ATC）[6]、电子病历基本架构与数据标准[7]、卫生信息基本数据集编制规范[8]，对院内采集汇聚的诊断、手术、药品、检验、检查等临床诊疗和医技相关数据进行标准化处理。通过自然语言的同义词表、医学术语的同义关联词表，在数据挖掘算法指导下对不同文字表达但含义相对的字段信息进行归一，为后续应用提供正确且统一的信息表达。（4）数据结构化。院内采集汇聚的数据中病历文书、护理记录、检查所见、检查结论等医疗数据存在较多非结构化或半结构化记录。大数据平台通过自然语言处理技术，结合医疗专业术语的语义结构，将医疗语义信息从原始自然语言表达扩展分析为结构化的 Key－Value 模式，对数据依据主题字段进行划分。数据治理针对电子病历、影像检查报告、手术记录、临床诊断等 10 类记录进行结构化处理，覆盖常规的检验、检查、症状、疾病生命体征、家族史、婚育史、过敏史、手术、病理、药品医嘱等临床字段数据，形成超过 7 000 条字段的数据集。

2.1.3 数据应用 数据采集加工后，建立多个应用充分挖掘数据价值。（1）病历全景视图/时间轴。将患者全病程数据按时间次序进行整合，形成全景视图。（2）病例检索和科研子系统。根据科研相关流程进行纳排人群、指标提取、统计分析等。（3）探索发现。通过数据可视化、桑基图等激发科研灵感。

2.2 数据质量评价模型构建

2.2.1 确定数据维度和定量测量方法 参考国家数据质量评价标准[9]并依据中肿实际情况，将数据质量问题划分为 5 个维度并确定相应定量计算方法，见表 1。

表 1 数据评价维度及定量测量计算方法

数据维度	指标名称	指标描述	计算方法
完整性	单字段完整性	单表单字段中，按照业务规则要求，数据元素有值的程度	$X=A/B$ A = 单字段元素有值的数量 B = 被评价的字段元素总数
	数据记录完整性	单表或多表中具有业务逻辑关联的多字段全部有值的程度	$X=A/B$ A = 多字段元素全部被附值的数量 B = 被评价的字段元素总数
规范性	格式合规性	数据类型、长度、精度等数据格式是否满足预期	$X=A/B$ A = 符合格式预期的数据元素数量 B = 被评价的字段元素总数
	数据重复率	特定字段、记录或数据集意外重复的比例	$X=A/B$ A = 意外重复的数据元素数量 B = 被评价的字段元素总数
准确性	内容正确性	数据内容是否符合预期	$X=A/B$ A = 符合内容、数值预期的数据元素数量 B = 被评价的字段元素总数
	数据唯一性	特定字段、记录或数据集唯一性	$X=A/B$ A = 非重复数据元素的数量 B = 被评价的字段元素总数
一致性	相同数据一致性	同一数据在不同位置应该保持一致，且内容的修改和删除应保持不同位置的同步	$X=A/B$ A = 满足要求的数据元素的个数 B = 被评价的数据元素总数
时效性	时效性	针对特定对象一系列状态相关变量的时序性逻辑，如一系列产前超声应先于妊娠结果	$X=A/B$ A = 满足要求的数据元素的个数 B = 被评价的数据元素总数

2.2.2 确定质量评价业务覆盖范围，梳理评价规则 根据院内信息系统情况，确定数据质量评价业务覆盖范围。以临床医疗数据为核心，范围涵盖 HIS、EMR、LIS、放射信息系统（Radiology Information System，RIS）、移动护理、手术麻醉、重症监护、病案系统、体检等临床业务信息系统，见表 2。依据国家数据质量评价标准[9]、数据管理能力成熟度评估模型[10]、国家电子病历规范[7]，并结合实践中常见问题情况和质控重点梳理评价规则，见表 3。

表 2 质量检查覆盖的业务系统/模块范围

系统通用名	系统子模块示例
HIS	患者（含门诊、住院）的基本信息、就诊情况、病历、诊断、医嘱、用药、耗材、手术、输血、检查、检验等信息
EMR	门诊患者的门诊病历，住院患者的入院病历、病程、术前讨论、术后情况、出院小结、会诊记录等全部文书
LIS	检查患者基本信息、身份信息、检查项目、检查细项、细项结果及正常值范围
RIS	包括但不限于病理信息管理系统、超声信息管理系统、电生理信息管理系统（心电）、内镜信息管理系统、骨穿信息管理系统、肺功能信息管理系统、核医学信息管理系统等的患者基本信息、身份信息、检查项目、检查方式、检查所见、检查结论等信息
移动护理系统	包括但不限于体征采集、医嘱执行、护理记录、出入量记录、护理评估、健康宣教、护理计划、不良事件上报、压疮评估、交班报告、标本采集、输血记录等信息

续表 2

手术麻醉信息管理系统	包括但不限于手术发生前后系统记录的术前访视、手术麻醉单、术后恢复、手术医嘱、术中护理、术中麻醉、手术申请、手术安排等信息
重症监护信息系统	包括但不限于患者信息的采集、存储、展现、包括重症监护记录单、管道管理、护理操作记录、评分记录、药品执行记录、给药记录、出入量记录、出入液体总量记录、转出转入记录等信息
病案管理系统	包括但不限于编目后的病案首页信息等
体检信息系统	患者基本信息、单位基本信息、体检项目清单、各项检查结果及正常值范围、各科室检查结论、终检结论、相关影像原始文件等

表 3 各数据维度规则举例

数据维度	规则举例
完整性	患者护理记录，血压数据中的收缩压和舒张压任意一个有值时血压观察时间都不能为空
规范性	门诊药品医嘱、同一次就诊、医嘱项目内容、医嘱开立时间都相同，怀疑重复
一致性	入院记录、入院西医诊断疾病名称、男性患者不应有妇科疾病、女性患者不应有男科疾病
准确性	病案费用、应收总费用应等于各子项之和
时效性	医嘱时间早于就诊时间或入院时间

2.3 针对各维度量化评分计算

为对数据质量情况进行直观判断，根据前述各规则判断的数据问题计算总分，规则如下：分维度计算分值：单维度分值 = ∑（1 - 问题数据量/数据总量）/规则数量。举例说明：规范性规则有两条，计算公式为：｛（1 - 规则 1 错误量/数据总量） + (1 - 规则 2 错误量/数据总量)｝/2。各维度加权计分：总分 = ∑权重 * 单维度规则分。

3 阶段性成果

3.1 建设以患者为中心的真实世界大数据平台

通过对 30 余个业务系统全覆盖式的数据采集、清洗和集成，建成以患者为中心、以唯一主索引为关联、覆盖诊疗全过程数据且 T + 0 实时更新的真实世界大数据平台，目前已汇集从 2000 年至今的 160 余万患者全量数据，治理病历数超 1 000 万，形成结构化字段 7 000 余条，并基于患者维度的全景数据将原始记录按照时序排列，形成患者诊疗时间轴。

3.2 建设数据质量监管系统

通过构建数据质量评价标准，按照原生系统数据质量的完整性、准确性、规范性、时效性、一致性 5 个维度，共设置 224 个参评规则，参评字段 158 个，可视化呈现数据质量情况。可以看出中肿数据在规范、时效、准确性等方面质量较好，而其一致性和完整性的质量稍差，大部分是历史原因所致的病案首页缺漏和不一致。

3.3 建立可溯源易反馈体系

针对每条参评规则，可查看规则设定的具体表、字段、错误的情况。为易于排查反馈，给出原生库结构化查询语言（Structured Query Language，SQL）排查语句和问题样例。最大程度上将检索出的问题以最便捷的形式反馈给相关原生系统，极大减少沟通时间成本。

4 成效与创新

4.1 实现数据质量评价定位问题的具象性、监控问题的持续性和实时性

数据质量评价的目的不仅是从宏观层面多维度了解当前数据质量情况，更重要的是方便发现问题，并有效反馈至原生系统进行数据整改。因此监管评价系统反馈数据具体情况在一定程度上减少了原生系统排查问题的工作量和难度。既往研究中的医疗数据质量管理系统功能多停留在问题发现层

面，并不直观呈现问题来源。而本研究所构建系统除了呈现发现问题的所属表和字段外，还通过下钻页列举问题样例，利用对应关系创新性地给出原生系统问题排查的 SQL 语句，以期以最易用、快捷的方式定位问题和成因，保障数据治理环节的流畅性，提高整改效率。同时数据问题的发现并不是横断面的，而是长期连续性的，通过持续实时监控可以得到初次排查整改的问题反馈，实时发现新产生问题，从而形成评价－反馈－治理的良性闭环。

4.2 大数据背景下实现跨信息系统的数据质量多维度评价

通过患者主索引将多个信息系统的数据关联起来，不仅实现单一系统的数据评价，还可以发现跨系统之间存在的数据问题，如时效性维度，即患者手术系统操作时间应在入院时间之后；一致性维度，即各系统姓名一致等。本研究通过各维度量化评分的计算将既往分散的规则付以权重，以百分制形式呈现，以便管理人员直观了解数据质量情况。利用大数据平台相关技术将各信息系统数据进行有效关联和分析，开展更全面的数据质量评价。

4.3 终末数据治理和初始监管并行，事后评价向事中监控过渡

通过人工智能技术对各系统数据进行清洗以构建大数据平台，利用 EMPI、NLP、标准化归一等技术手段，对原始数据进行终末治理，提升数据可用性，为各智能化应用场景提供坚实的数据基础。同时在监管过程中核查数据情况，记录并呈现数据问题。通过发现事后问题指导并协助原生系统进行相关功能优化，同时在事中进行实时监控，将问题发现的节点前置，避免问题数据产生，形成监控－反馈－治理的良性闭环，不断提升数据治理水平。

5 结语

本实践通过对院内各业务系统的数据采集、清洗和整合构建统一的大数据平台。在此基础上结合国家、行业标准实现对院内数据的终末治理、系统质量评价，并通过构建实时数据质量监管系统实现数据问题可视化、易溯源，推动形成监控－反馈－治理的良性闭环，发挥治理工具价值并实现数据增值。受人为因素影响，如操作不规范、流程不合理可能导致数据缺失，跨系统对接不流畅可能导致数据不一致，因而数据治理实质是管理。只有将质量评价中发现的问题进行深度剖析，协调人力落实问题整改、优化流程、规范业务，才能真正为提升数据质量奠定基础。后续将持续优化数据评价模型、建立完善管理体系，全方位提升院内数据治理能力，提高数据的可信性、可用性。

参考文献

1 费晓璐，李嘉，黄跃，等. 医疗大数据应用中的数据治理实践［J］. 中国卫生信息管理杂志，2018，15（5）：554－558.

2 张振，周毅，杜守洪，等. 医疗大数据及其面临的机遇与挑战［J］. 医学信息学杂志，2014，35（6）：2－8.

3 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.

4 袁莎，沈丽宁. 医疗数据质量评价指标体系研究［J］. 卫生软科学，2020，34（12）：12－16.

5 熊兴江，马敬东，徐承中. 电子健康档案数据质量评估与治理的综述研究［J］. 中国卫生信息管理杂志，2018，15（6）：637－642.

6 世界卫生组织药物统计方法整合中心. 药品解剖学、治疗学及化学分类系统（Anatomical Therapeutic Chemical，ATC）［M］. 北京：中国协和医科大学出版社，2003.

7 原卫生部，国家中医药管理局. 电子病历基本架构与数据标准（试行）［EB/OL］.［2022－06－14］. http://www.nhc.gov.cn/bgt/s6718/200912/45414.shtml.

8 原卫生部. 卫生信息基本数据集编制规范 WS/T 370－2012［EB/OL］.［2022－06－15］. http://wsbz.nhc.gov.cn/wsbzw/article/10/2020/10/40288ad475307e2f017553770184007a.html.

9 中国国家标准化管理委员会. 信息技术数据质量评价指标 GB/T36344－2018［EB/OL］.［2022－06－15］. https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=D12140EDFD3967960F51BD1A05645FE7.

10 中国国家标准化管理委员会. 数据管理能力成熟度评估模型 GB/T 36073－2018［EB/OL］.［2022－06－15］. https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=B282A7BD34CAA6E2D742E0CAB7587DBC.