医疗数据治理——构建高质量医疗大数据智能分析数据基础

阮彤，邱加辉，张知行，叶琪
华东理工大学计算机科学与技术系，上海 200237

摘要
以专病真实世界研究为背景，分析了医疗数据治理和数据可用性存在的各种问题。从医院数据治理出发，延伸到区域数据治理、专科联盟数据治理以及医疗标注数据与知识型数据的治理，总结了不同层次、不同层面数据治理的共性方法与个性方法。在此基础上，进一步探讨了数据治理中的主数据管理、元数据管理、数据质量控制等关键技术和关键环节。最后，给出了医疗大数据标准的基本框架，并基于现有的数据治理评估标准，对医疗大数据治理的现状进行了评估。

关键词
医疗数据治理；数据可用性；元数据；主数据

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019002

Medical data governance: building the data foundation for intelligent analysis of high quality medical big data

RUAN Tong, QIU Jiahui, ZHANG Zhixing, YE Qi
Department of Computer Science & Engineering, East China University of Science and Technology, Shanghai 200237, China

Abstract
The various problems of medical data governance and data availability in the context of real-world research on a specific disease type were analyzed. The definitions and concepts of the data governance were provided, such as hospital data governance, regional data governance, data governance of disease-specific alliances, governance of medical annotation data and knowledge-based data. Here, the common methodologies and personality methodologies of data governance were summarized. Furthermore, the technologies of master data management, metadata management and data quality control in data governance were further discussed. Finally, the basic framework of medical big data standards was given, and the medical data quality was simply evaluated which based on the existing data governance evaluation standards.

Key words
medical data governance, data availability, metadata, master data

1 健康医疗数据治理问题背景与发展现状

医疗健康大数据与人工智能呈蓬勃发展的态势。一方面，医院互联互通、国家与省市大数据中心的建设，为医疗健康大数据奠定了良好的基础；另一方面，个性化诊断、疾病预测与辅助决策支持系统等各类医疗人工智能应用也在不断涌现。

国家对发展医疗大数据和人工智能非常重视，近几年颁布了一系列指导意见和措施。国办发〔2016〕47号《国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见》提出建设全民健康保障信息化工程一期项目。作为该项目的前期工程，2015年国家卫生和计划生育委员会统计信息中心启动了十省互联互通项目，开展相关技术的验证工作。截至2016年，该项目已经接入上海、湖南、湖北、江苏、浙江、福建、重庆、内蒙古、辽宁、北京10个省级健康医疗大数据平台，目前已完成所有省级健康医疗大数据平台的接入。为深入贯彻落实国办发〔2016〕47号文件精神，加快国家健康医疗大数据中心及产业园建设试点项目对接工作，2016年10月，福建、江苏以及福州、厦门、南京、常州被确定为健康医疗大数据中心与产业园建设国家试点工程第一批试点省市。2017年12月，健康医疗大数据中心第二批国家试点启动。根据部署，健康医疗大数据中心第二批国家试点已在山东、安徽、贵州3个省开展，并与第一批试点的江苏、福建一起作为东、南、西、北、中5个健康医疗大数据区域中心建设及互联互通试点省。目前，医疗领域已经积累了大规模临床数据，并有了国家级别的数据采集和汇聚方案，保障了持续的数据增加和更新，并且为人工智能应用提供了基础。

然而，在利用前期基于医院或国家卫生健康委员会的区域平台数据进行临床科研和人工智能应用开发的过程中，即使在病人数量足够的情况下，数据的可用性依然存在问题。这里既有数据本身的问题，也有数据流程管理问题和数据使用权限问题[1]。例如，若研究特定治疗方案对心衰与大肠癌的影响，需要从电子病历中分别构建心衰队列和大肠癌队列。对于心衰队列，入组条件是电子健康档案中患有心衰疾病的病人，控制变量是是否吃了与心衰治疗相关的中药，终点事件是180天再入院率。对于大肠癌队列，入组条件是电子健康档案中患有大肠癌的病人，控制变量是是否手术，终点事件是复发或3~5年生存期。在构建队列的过程中存在以下问题。

- 数据完整性不够。例如，诊断心衰需要心脏彩超数据，诊断大肠癌需要病理和死亡数据。对于医院来说，心脏彩超数据和病理数据是有的，但这些数据被分散在不同的检查系统中，需要将这些数据集成在临床专病库中。此外，与终点事件相关的死亡信息和再入院率数据存在缺失的情况。如果病人不是在医院死亡的，系统就无法知道病人的生存状态，因此死亡信息无法被获取。病人也有可能到其他医院就诊，这样就无法统计再入院率这个数据了。如果是区域平台，各省只有检验数据，没有检查数据。
- 数据精准度不够。例如，心衰与心功能分级有关，而肿瘤需要分期，在电子健康档案数据里面，很多病人的心功能分级数据或肿瘤分期数据是没有的。
- 数据一致性不够。以患者基本信息和时间信息为例，由于填错或者其他原因，患者的基本信息在不同系统中可能是不一样的，而不同系统的时间格式也可能是不一致的。

● 数据准确性不够。病人的很多ICD编码不在国家卫生健康委员会编码规范里，此外，很多疾病名称与ICD编码系统中的疾病名称不一致。

综上所述，随着医疗大数据与人工智能技术的深入发展，数据可用性已经成为制约医疗健康大数据智能分析应用发展的瓶颈。数据存在各种各样的问题，具体如下。

● 原始数据在录入过程中有数据错漏、数据不完整等问题。

● 由于缺乏统一的元数据标准，数据融合困难。

● 由于缺乏统一的主数据管理，病人、医生等医疗应用中的核心数据实体难以被唯一标识并实时更新。

● 数据清洗缺乏统一的策略，导致数据被多次清洗，使用代价高。

● 由于缺乏元数据和主数据标准，即使数据被勉强放在一起，数据可达性也很差，无法知晓每个字段的确切含义和具体取值范围，难以基于简单的查询找到需要的数据。

● 大量医疗数据以文本、影像、图像等非结构化的方式存储，增加了管理和整合的难度。

另外，无论是在规划层面还是在操作层面，数据隐私管理、数据使用的权限与流程都缺乏指导性的技术标准和规范，由此导致虽然采集、存储了很多数据，但不知道谁可以用、应采用什么样的方法用。

总体来说，要让医疗大数据应用生根发芽，让数据驱动的医疗人工智能技术与应用落地，就必须从源头解决数据质量问题，解决在什么样的情况下可以用什么样的数据的问题，解决数据录入、采集、融合与使用过程中多个环节的数据监督与控制问题[2-3]，这些问题都属于医疗数据治理问题。

2 医疗数据治理的概念、分类与阶段

2.1 医疗数据治理的内涵与外延

维基百科将数据治理定义为组织机构遵循的一个既定的流程，以确保整个生命周期中的数据质量。国际数据管理协会（Global Data Management Community，DAMA）在《DAMA数据管理知识体系指南》中将数据治理定义为数据资产管理的权威性和控制性活动，是对数据管理的高层计划与控制，包括在数据管理和使用层面进行规划、监督和强制执行[4]。我国电子工业标准化技术协会信息技术服务分会（Information Technology Service Standards，ITSS）基础标准工作组认为，数据治理包括：对信息利益相关者的需要评估；确保有效助力业务的决策机制和方向；确保对合规和绩效的监督[5-6]。

根据上述定义可知，数据治理需要一个权威的组织机构，这个组织机构需要对数据资产制定权威性的计划，并控制活动。对于企业内部的数据治理，企业可以建立自己的数据治理机构，对数据行使治理的权力[7]。但是，医疗数据相对复杂和分散。一方面，医院、国家卫生健康委员会、体检机构、医疗科研、保险机构、互联网健康机构以及病人都存有数据；另一方面，医疗数据所有权难以明确，数据治理的权威机构缺乏规定[8]。

为此，国家卫生健康委员会在2018年9月研究制定了《国家健康医疗大数据标准、安全和服务管理办法（试行）》（以下简称《管理办法》）。在《管理办法》的第二条中规定：我国公民在中华人民共和国境内所产生的健康和医疗数据，国家在保障公民知情权、使用权和个人隐私的基础

上，根据国家战略安全和人民群众生命安全需要，加以规范管理和开发利用。《管理办法》明确了国家和政府机构对数据的使用权力，指明了数据治理的权威机构是国家卫生健康委员会，国家卫生健康委员会和相关部门负责统筹规划、指导、评估、监督全国健康医疗大数据的标准管理、安全管理和服务管理工作。县级以上卫生健康行政部门和相关部门负责本行政区域内健康医疗大数据的管理工作，是本行政区域内健康医疗大数据安全和应用管理的监管单位。各级各类医疗卫生机构和相关企事业单位是健康医疗大数据安全和应用管理的责任单位。

2.2 医疗数据治理的分类

本文的数据治理主要围绕医院和区域卫生管理机构拥有的数据展开。将医疗数据治理按管理机构分为3类：一是医院数据治理；二是区域数据治理；三是专科联盟/专科医联体/专病中心的数据治理。此外，还有医疗标注数据与知识型数据治理。

2.2.1 医院数据治理

医院数据治理被定义为：医院对其数据资产的管理和控制，支撑并保障数据被安全、高效地交换与使用。**图1**展示了医院数据治理体系。

（1）组织与人员机构管理

医院需要成立专门的数据管理部门，完成流程和规范制订、数据质量保证和质量控制、流程审批等工作，并对数据使用方和IT设施建设方进行管理。目前，诸多医院已经专门成立了大数据部门承担这项工作，也有医院将这项工作放在信息科或

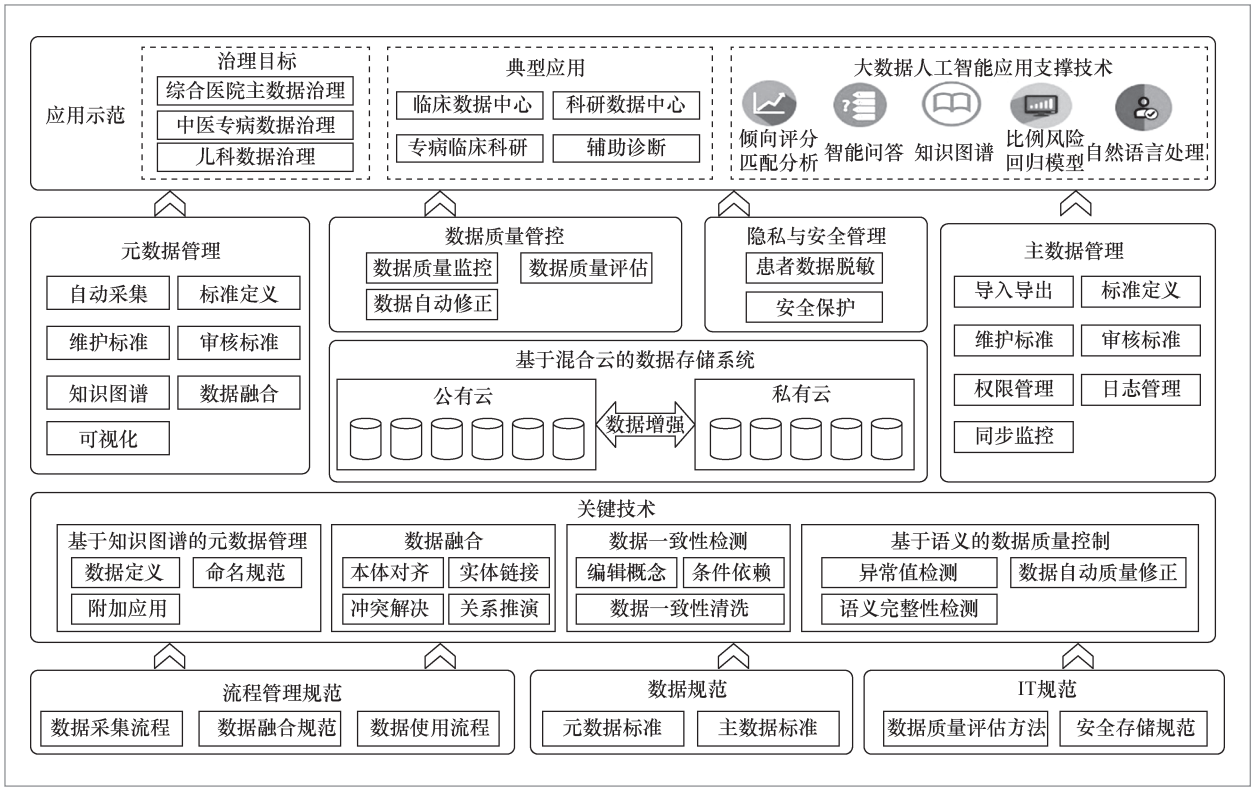

图1 医院数据治理体系

病案管理室。

（2）规范管理

规范管理包括流程管理与支撑规范、信息规范与数据规范。流程管理规范规定了何人在何种应用场景下，通过何人的审批可以操作何种类型的数据。首先，流程管理规范规定了人员或是角色的配置与管理，确定数据标准、更新维护、数据质量跟踪等内容由谁完成。其次，规定了各种审批流程。

相比数据采集、加工与存储的流程，数据使用流程的制度管理更为重要。部分医院依据确定的审批流程管理数据的使用，临床科室可以作为数据使用方提交数据使用要求，大数据管理部门以及其他数据管理方进行审批。但是，目前审批过程缺乏详细的附加规范，比如何种职责的科室和医生在什么情形下可以导出数据、数据能否离院、应由哪一级来审批决定，这些都需要具体规定。除管理规范外，不同角色的人员如何协作互动完成既定的工作，也可通过制定流程支撑规范实现。比如，科研流程不是简单的数据导出和使用过程，通常可能需要多次的数据字段确认和细化、数据清洗和融合，因此，需要面向特定任务建立流程支撑规范，确保信息科、数据管理部门以及临床医生的有效协同。

信息规范包含隐私、数据权限管控规范和质量评估规范等。数据规范包括不同业务系统，如电子病历数据规范、医院信息系统数据规范。对于数据整合过程和整合后的系统来说，需要主数据和元数据规范以及相关的数据质量规范。

以上规范的制定有助于完成以下目标：数据有明确和准确的定义；数据有明确的责任方；数据有清晰的存储方式与合理的时间期限；数据加工方法明晰；数据访问方式与控制明确；数据内容符合标准要求与质量要求。

（3）平台、工具与关键技术

数据治理需要有相应的平台与工具支持，数据清洗过程需要抽取—转换—加载（extract-transform-load，ETL）工具[9]、文本结构化工具[10]等。大数据管理需要支持多种存储机制，管理平台需要包含元数据管理、主数据管理、数据权限管理、审批流程管理、数据质量评估、数据质量监控等工具。此外，数据的使用需要各种统计模型和人工智能算法。

（4）数据智能化分析应用

整个数据治理应该是应用驱动的。治理可以围绕临床科研、医院管理、辅助诊疗以及各种人工智能应用展开。

2.2.2 区域数据治理

区域数据治理与医院数据管理内容相似，但实施起来难度更高，具体体现在以下几点。

（1）主数据管理和元数据管理的复杂度高

病人基础数据是临床医疗信息的主数据。区域数据来源于多家医院，每家医院病人用的身份标识不一样，病人基础信息也会有差异。需要通过统一标识来统一病人的主数据，并关联病人在不同医院的就诊记录。另外，每家医院的健康档案的数据结构、疾病、检验、症状的名称也有差异。如果希望做统一元数据管理，不仅涉及医疗信息系统设计问题，也涉及医学语言的表达能力以及专科之间的差异问题，目前还没有哪个区域能比较好地解决这个问题。

（2）数据安全性管理更严格，审批和流程管理更复杂

由于区域数据量比较大，病人的就诊数据在时序上更完整，因此数据泄露带来的严重性更大，区域对数据安全管理的要

求更严格。另外，区域卫生管理机构只是行使数据管理权，数据使用者很多是医疗机构，在管理机构对医疗机构的数据使用申请进行约束和审批的过程中，由于无法了解医疗机构的真实意图，因此需要通过更严格的审批和管理流程，保障数据的合理与安全使用。

2.2.3 专科联盟/专科医联体/专病中心的数据治理

专科联盟/专科医联体/专病中心的数据治理问题更复杂。专科联盟一般由权威医疗机构牵头，但是其牵头单位并没有行政权力，联盟单位之间的协作共享完全是一种自愿的行为。因此，专科联盟形式的医联体除了要解决区域医联体中碰到的技术问题外，还要解决数据共享后的利益分享问题，确保医联体每个成员能在数据共享活动中受益。这项工作有部分医疗信息厂商正在做探索性的研究，主要通过区块链确立数据来源，并通过智能合约实现临床科研上利益的共享。专病中心可以看作专科医联体建立的数据中心，数据中心需要解决医联体成员对专病信息汇聚的技术问题。为了形成高质量的数据中心，部分专病中心尝试采用了制定复杂的数据入库规则、定时数据获取制度、规范化科研病例报告表单定义等措施。

2.2.4 医疗标注数据与知识型数据治理

除了上述3类管理机构的数据治理问题之外，还有医疗标注数据与知识型数据的治理问题。数据治理主要面向的对象是病人数据，但在医院协作共享过程中，知识型数据[11]也必不可少。在面向人工智能的应用时，需要大量的标注数据，这些数据的管理和利用也应该属于数据治理的范畴。

标注数据主要是针对电子病历文本、影像等非结构化数据进行实体、属性、关系等标注得到的数据，标注数据的质量对训练深度学习或神经网络模型起着决定性作用。为了实现对标注数据的治理，应该针对不同粒度的实体建立一套完整的标注规范，对标注过程的各要素进行规范化管理，并对标注结果进行交叉验证等。

由于不同机构对知识型数据的表示各不相同，且知识之间存在复杂的关系，因此需要对知识型数据进行统一的表示，明确知识型数据之间的关系。此外，知识还会不断地增长，需要建立知识管理机制，确保知识型数据能够动态地扩充。

3 医疗数据治理工具平台与关键技术

建立临床数据治理需要平台和工具，也要攻克一系列的关键技术。医疗数据治理工具平台应包含数据存储子系统、元数据管理子系统、主数据管理子系统、数据质量管控子系统以及患者数据脱敏工具等。为了构建这样的工具平台，需探讨知识图谱、数据融合算法、数据自动一致性检测等多种技术。本文主要探讨与医疗数据关系最密切的3项技术：元数据管理、主数据管理以及数据质量监控与管理。

3.1 元数据管理

目前医院信息系统中存在数据模式描述文档不全、系统之间数据关联不清晰、系统值域标准不统一等问题，这对数据的集成造成了极大的困扰。在区域层面，这些问题更严重。因此，需要通过元数据管理获取业务系统中数据的含义，辅助数据理解，增加分析的敏捷性。元数据管理可以提高数据的可访问性、一

致性及可用性，为多种来源数据的整合搭建了桥梁[12]。

相比利用传统字典表以及数据字段名称的元数据初级表示方法，完整的元数据管理有完整的字段定义、与数据源的对应关系以及不同数据来源元数据之间的映射关系。

具体而言，元数据管理子系统主要由采集层、标准层、分析层和应用接口层组成。采集层从各种医疗信息系统内获取元数据，查看元数据的变化和更新。标准层保存了元数据的标识信息、内容信息与模式信息等。由于医院信息系统各异，标准层实现了将元数据映射到标准集合以及将不同的元数据进行互操作的功能。分析层主要提供了对元数据的管理、分析与查询。元数据库与数据源存在对应关系，当用户通过统一入口提交查询服务时，可以根据元数据库提供的特征找到对应的信息资源，重组之后呈现给用户。因此，应用接口层除了提供元数据访问的限制和保护外，同时还服务于各个应用程序。

与其他领域相比，医疗领域的元数据规范相对比较成熟，如原卫生部颁布的《国家卫生计生委办公厅关于印发住院病案首页数据填写质量规范（暂行）和住院病案首页数据质量管理与控制指标（2016版）的通知》（国卫办医发〔2016〕24号）、《病历书写规范》（卫医政发〔2010〕11号）、《电子病历基本规范》（卫医政发〔2010〕24号）、《卫生信息基本数据集编制规范》（WS 370-2012）、《卫生管理基本数据集》（WS374-2012）与《电子病历基本架构与数据标准》（卫办发〔2009〕130号）等。在数据值编码标准方面，国际上有疾病分类编码ICD-10、手术操作编码ICD-9以及SNOMED术语库，国内有国家标准《卫生机构（组织）分类与代码表》（WS218-2002）、《社会保险药品分类与代码》（LD/T90-2012）和《中医病证分类与代码》（GB/T15657-1995）。

然而，在使用过程中，这些标准会根据应用进行不同程度的删减和扩充，甚至出现错误的使用。因此，基于标准建立一个元数据管理机制，可方便地在标准上扩充，并可以关联不同的应用。从技术角度来说，这个管理机制可以基于知识图谱技术建立。图谱可以方便地给出元数据的定义，包括概念、概念层次、属性、属性值类型、关系、关系定义域概念集以及关系值域概念集，并且可以通过添加规则或公理来表示模式层更复杂的约束关系。借助知识图谱图结构的表达能力，可以表达关联关系、同义关系、上下位关系与实例关系等，从而建立更加灵活的数据约束条件[13]，从而方便地表达和扩充元数据。借助于图谱已有的模式对齐、实体匹配与冲突检测算法，可以在语义层次对图谱进行维护，在此基础上实现自动的数据融合算法。**图2**给出了疾病“心力衰竭”的部分知识图谱，包括该疾病名称的同义词：心功能不全、心衰、心脏衰竭、心力衰竭等，所属部位是“心脏结构”。

利用该图谱还能较好地实现电子病历的标准化。如**图3**所示，将电子病历中不规范的文本进行标准化，例如将疾病名称“心衰”标准化为“心功能不全”或“心力衰竭”，将“脑梗死”标准化为“脑梗塞”，将药物“异舒吉”映射到标准药物名“硝酸异山梨酯”等。

3.2 主数据管理

医疗数据的主数据主要有病人信息和医生信息两类。本文以病人信息为例，说明主数据管理的问题和难点。目前，在医院层面，各业务系统对病人的信息分别进行存储，但大型医院都建立了临床数据中心

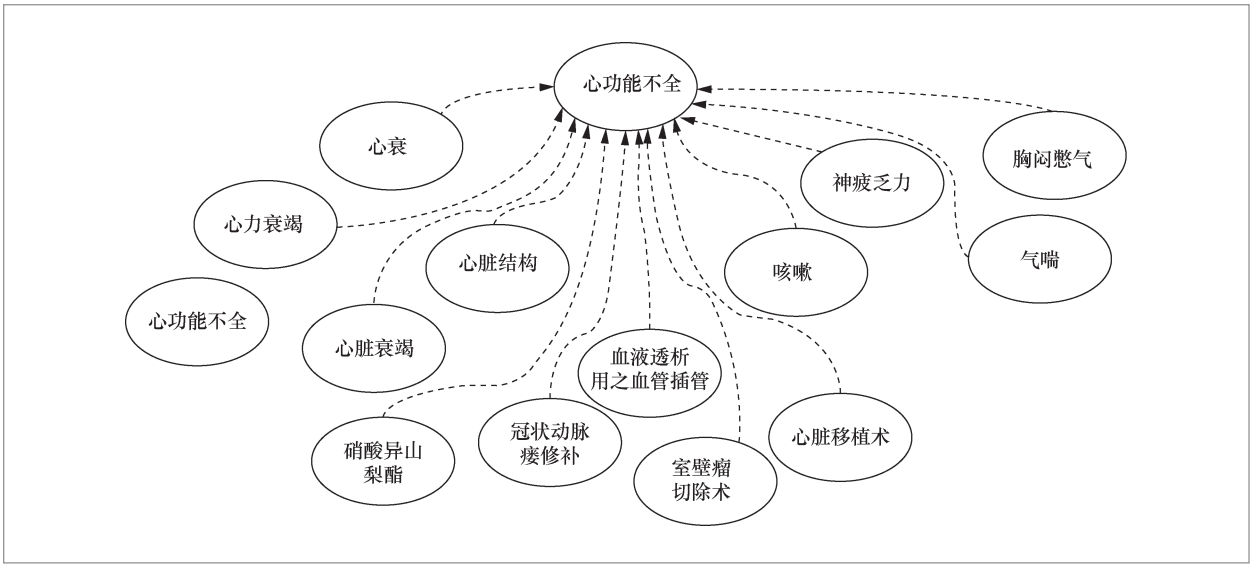

图 2 疾病“心力衰竭”的部分知识图谱

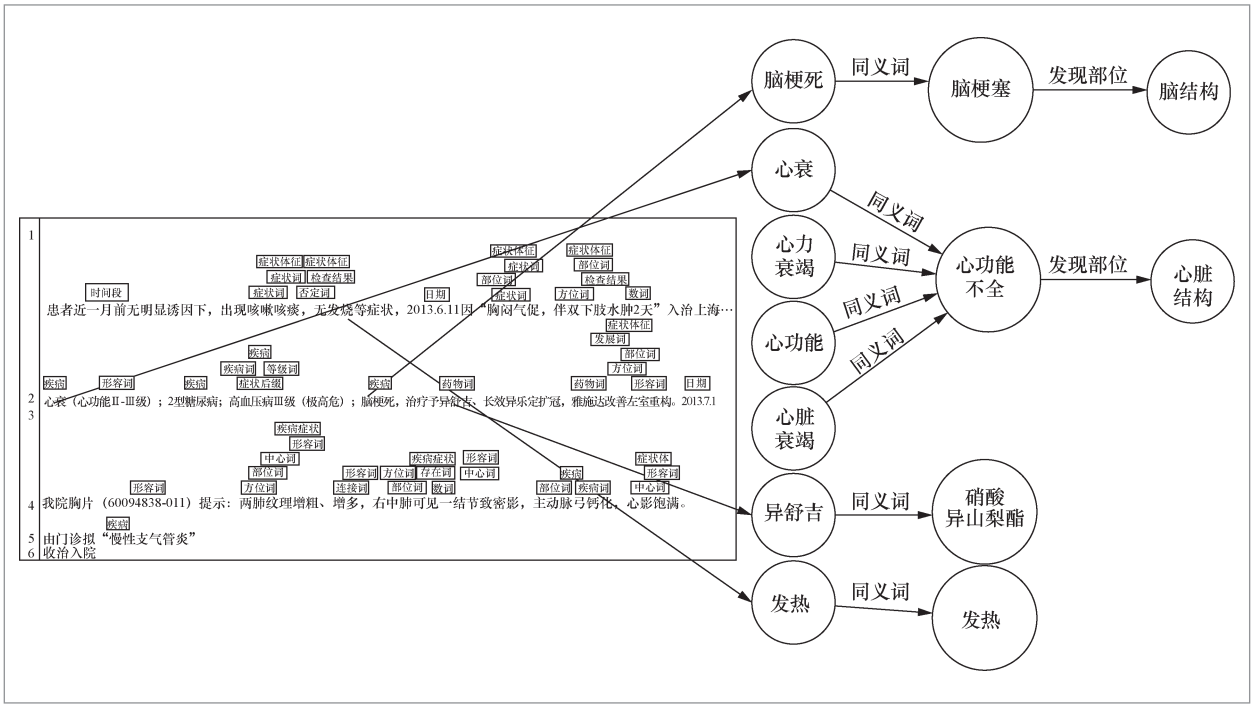

图 3 利用知识图谱实现电子病历的标准化

（clinical data repository，CDR），为了唯一标识一个病人，需要通过构建病人主索引号（enterprise master patient index，EMPI）将存储于不同系统的病人关联在一起。这里有两个问题需要解决。第一，如何构建EMPI。识别不同系统中同一个病人不同ID之间的映射关系十分困难，特别是在区域平台上每个系统都有独立的ID，导

致这个问题更复杂了。虽然这个问题可以通过匹配医保卡号部分解决，但由于有些病人没有医保卡，有医保卡的也不是每次都用医保卡，因此，要求病人通过身份证和手机号进行实名认证是必要的。第二，一个病人的基础信息（如年龄、性别等）可能同时存在于医院信息系统（hospital information system，HIS）、实验室信息管理系统（laboratory information management system，LIS）和影像存储与传输系统（picture archiving and communication system，PACS）等系统中。而各系统的侧重点不同，难免会造成数据填写质量不一致或数据未及时更新等问题。

为此，需要在定义系统主数据的情况下，构建主数据管理中央库，解决主数据碎片问题。可以从各业务系统抽取数据，并进行数据融合，形成完备的主数据信息，然后再将主数据信息分发给各业务系统，保证各业务系统中这些信息的准确性和完整性。这样就形成了公共的重要属性由主数据管理系统管理、各业务系统的特色属性由各系统独立管理的模式[14]。

在构建主数据管理库时，首先需要从多个异构的业务子系统中以ETL的方式抽取关键数据，然后，利用元数据库对其中的编码、描述进行标准化[15]。接着，由于多个业务系统的数据可能不一致，还需要通过匹配算法完成对数据的错误消除和信息融合。对于匹配不到的孤立信息，要加以监控跟踪，进行人工处理。同时，以增量学习的方式不断改进匹配算法。最后，将归整好的主数据信息存入主数据库。

以病人信息为例，病人的出生年月、性别等信息在各系统中都有，但由于HIS是从身份证读取的信息，因而相比其他系统采用人工录入的方式来说更准确。然而，病人的血型信息通常在LIS中是更准确的，甚至可能在门诊/急诊工作站中也没有这些信息。通过主数据管理系统，可以从各系统中分别获取信息，根据可靠性、一致率等进行信息校验融合，最后形成该病人的完备信息，然后再将完备信息下发到各业务系统。

在这些主数据管理的基础服务之上，还可以提供诸如患者画像等高级功能。目前患者就诊时，医生仅能够通过患者自述的方式快速了解慢性病史、过敏史等信息。尽管可以查阅历史档案，但对于医生来说很不方便。而通过对诊断等数据的分析，可以形成患者慢性病、过敏史的标注信息，将这些信息合并到病人的主数据中形成患者画像。患者就诊时医生可以直接从医生工作站上读取到这些标注信息，从而准确、快速地获取这些必要信息。此外，在多家医院之间实现主数据共同管理之后，还可以打通医院之间的信息壁垒。若患者在A医院诊断出患有某慢性病，当他下次去B医院就诊时，医生也能快速得知该患者的这个信息，从而使就诊更高效、更准确。

3.3 数据质量管控子系统

从数据产生过程来看，医疗数据质量问题主要来源于3个方面。

一是原始信息采集有误差。在医疗系统内数据采集主要通过手工方式录入，在医生或护士输入信息的过程中，可能会有意或无意地将数据错误引入系统。

二是数据融合过程发生问题。在对不同来源的数据进行融合时，数据格式和语义可能会有误差或不一致，导致融合结果有错。

三是与数据的应用场景不匹配[16]。例如，如果要进行病例统计，现有临床电子病历数据就能满足统计场景的需求。但如果

要做大肠癌疗效分析，现有临床电子病历数据就难以满足分析场景的要求，还需补充病理数据。

因此，在医疗数据治理流程中，需要了解最终的使用场景，也需要从业务系统的数据源头控制质量，并保证每个融合和加工过程的正确性。另外，当发现错误的时候，可以实现自动或半自动的修正。因此，质量管控平台包括了数据质量实时监控、数据质量后评估以及数据的自动修正。数据质量实时监控主要针对从业务系统抽取的或是从外部传送的接口数据，通常从及时性、有效性和完整性等几个指标监测接口内容本身的数据质量问题，还需要对采集程序进行监控，如接口采集程序是否正常启动、是否正常结束等。数据质量后评估是指对融合后的数据进行质量评估。首先从确定评价对象和范围着手，然后选取数据质量维度及评价标准，确定质量测度及评价方法，之后按照配置的评估指标执行评估，产生权重化的评估结果，最后生成质量结果和报告。数据的自动修正是指对于有错误或不一致的数据，部分数据可以进行自动化的探测和更正。例如，时间的表达可以是DD/MM/YY，也可以是YY/MM/DD，可以通过智能算法探测原始系统的表达方式，制定映射规则，实现日期格式的一致性。

除了上述内容之外，大数据存储的便利性与查询速度、医疗数据脱敏、流程管理工具、数据访问权限管理、数据的防复制等也是非常重要的问题[17]，此处不再详述。

4 医疗大数据治理标准与医疗数据治理能力评估

在单独的业务系统以及数据互联互通方面，国家卫生健康委员会已建立了相应的标准。但是，在医疗大数据方面，还缺乏专门的标准。《国家健康医疗大数据标准、安全和服务管理办法（试行）》指定，国家卫生健康委员将负责全国健康医疗大数据标准的制定工作，省级卫生健康行政部门将负责大数据标准的落地工作。从整体规范体系来看，需要制定的标准众多，应该包含**图1**中的流程标准、数据标准以及信息标准。需要特别指出的是，数据标准包含病人、医生、医院等主数据标准，也包含检查、疾病、用药等数据值标准。数据标准还需要对不同的医疗行为（如就诊、诊断、治疗等）进行标准化的描述。为了便于人工智能应用，对于文本数据和影像数据，也应有相应的标注规范。由于不同专科需要的数据字段有差别，有可能会进一步制定专科规范。总而言之，标准体系包含了相互关联的多个标准，覆盖面广，种类众多，标准的建设不是一蹴而就的。

除了医疗大数据治理标准外，值得一提的是医疗数据治理能力评估规范。数据治理能力评估规范可以考察企业对数据的管理能力，通过对企业的评估，可以逐步提高企业的数据治理能力。该评估标准经过修改后也可以用在医疗领域。

2010年，IBM公司发布的《IBM数据治理统一流程》描述了企业数据能力成熟度评价模型，将数据能力分为5个等级和11个功能域。2014年，能力成熟度模型整合（capability maturity model integration，CMMI）协会发布了企业数据管理成熟度（data management maturity，DMM）模型，该模型规定了6个职能域和25个过程域。

国内最早启动相关研究的机构是全国信息技术标准化技术委员会，该机构于2014年开始开展相关的研究，并在2018年3月15日颁布了国家标准《数据管理能力成

熟度评估模型》(GB/T 36073-2018)。该标准可以用于评估不同行业的企业的数据管理制度、手段、方法以及相关能力。标准从8个关键过程域考察了企业管理数据的能力。过程域和过程项包括：数据战略(数据战略框架、数据战略实施、数据战略评估、数据任务效益评估)、数据治理(数据治理组织、数据制度建设、数据治理沟通)、数据架构(组织数据模型、数据分布与整合、数据共享与应用服务、元数据管理)、数据应用(数据分析、数据开放共享、数据服务)、数据安全(数据安全策略、数据安全保护、数据安全审计)、数据质量(数据质量需求、数据质量检查、数据质量分析、数据质量提升)、数据标准(业务术语、参考数据和主数据、数据元、指标数据)、数据生命周期(数据需求、数据设计与开发、数据运维、数据退役)。评价方法主要采用评价问卷和调研访谈两种。评价问卷是根据数据能力成熟度模型定义每个域的评价指标，并对各个指标进行加权平均，计算自评结果。调研访谈则是根据重点问题，对相关单位人员进行输入了解，准备评估数据能力的真实情况。

上述成熟度评价模型都将成熟度定义为5个等级。以国家标准GB/T 36073-2018为例，5个等级被定义为：初始级、受管理级、稳健级、量化管理级和优化级。其中，初始级是指对数据需求的管理主要在项目级体现，没有统一的管理流程，是一种被动的管理；受管理级是指组织已经意识到数据是资产，根据管理策略的要求制定了管理流程，指定了相关人员进行初步管理；稳健级是指数据已经被当作实现组织绩效目标的重要资产，在组织层面制定了系列标准化管理流程，促进数据管理的规范化；量化管理级是指数据被认为是获取竞争优势的重要资源，数据管理的效率能够被量化分析和监控；优化级是指数据被认为是组织生存的基础，相关管理流程能够实时优化，能够在行业内进行最佳实践的分享。

参照国家标准对成熟度评价的5个等级，结合医疗数据管理的现状，对医疗行业的数据治理能力进行分析和评估后发现，目前大多数医疗机构的数据治理能力至多在受管理级，即把数据当作重要资产，进行了初步的数据集成工作，制定了一定的管理流程。但是，还远远未达到稳健级，目前的数据管理流程还远远没有达到标准化阶段，无法满足数据管理及应用能结合组织的业务战略需求、经营管理需求以及监管需求，甚至很多医院并没有人专门管理数据。

5 结束语

医疗数据治理是一个“修高速公路”的过程，人工智能和大数据挖掘应用是“跑车”。修路是一个基础设施建设，需要大量的投入，而且不会快速产生效果。但是，若没有成功的数据治理，数据可用性就会有问题，现有的人工智能和大数据挖掘算法的效率和有效性就会打折扣。因此，各级医疗机构和医疗管理部门应投入人力物力，建立有效的数据治理机制，保障国家人工智能和大数据战略的有效实施。

参考文献:

[1] LIN J C, FAN C T, LIAO C C, et al. Taiwan biobank: making cross-database convergence possible in the big data era[J].Journal Citation Reports, 2018, 7(1): 1-4.

[2] DEMARQUET G. Five key reasons

enterprise data governance matters to finance and seven best practices to get you there[J]. Journal of Corporate Accounting & Finance, 2016, 27(2): 47-51.

[3] 苏玉娟. 大数据技术与高新技术企业数据治理创新——以太原高新区为例[J]. 科技进步与对策, 2016, 33(6): 47-52.
SU Y J. Big data technology and the high and new technology enterprises innovate Data management: Taiyuan High-Tech Zone as an example[J]. Science & Technology Progress and Policy, 2016, 33(6): 47-52.

[4] SUSAN E. The DAMA guide to the data management body of knowledge (DAMA-DMBOK)[M]. Bradley Beach: Technics Publications LLC, 2010: 58-72.

[5] 张明英, 潘蓉. 数据治理白皮书国际标准研究报告要点解读[J]. 信息技术与标准化, 2015(6): 54-57.
ZHANG M Y, PAN R. Main points elaboration of international standard technical report "white paper of data covernance" [J]. Information Technology and Standardization, 2015(6): 54-57.

[6] 常朝娣, 陈敏. 大数据时代医疗健康数据治理方法研究[J]. 中国数字医学, 2016, 11(9): 2-5.
CHANG Z D, CHEN M. Research on governance methods of healthcare resources in big data era[J]. China Digital Medicine, 2016, 11(9): 2-5.

[7] BEGG C, CAIRA T. Exploring the SME quandary: data governance in practice in the small to medium-sized enterprise sector[J]. The Electronic Journal Information Systems Evaluation, 2012, 15(1): 3-13.

[8] 王融. 关于大数据交易核心法律问题——数据所有权的探讨[J]. 大数据, 2015, 1(2): 49-55.
WANG R. Discussion in the legal core question of the data ownership in big data trade[J]. Big Data Research, 2015, 1(2): 49-55.

[9] SEBASTIAN M, FELIX K, DENNIS T, et al. Ontology-based data integration between clinical and research systems[J]. Public Library of Science ONE, 2015, 10(1): 1-20.

[10] HANNELE H, KAIJA S, RIIKKA V, et al. Impacts of structuring the electronic health record: a systematic review protocol and results of previous reviews[J]. International Journal of Medical Informatics, 2014, 83(3): 159-169.

[11] NG P M L, LO M F, CHOY E. Improving China's corporate governance within the big data era-integration of knowledge management and data governance[C]// International Conference on Intellectual Capital and Knowledge Management and Organisational Learning, December 5-6, 2019, Sydney, Australia. [S.l:s.n], 2015: 183.

[12] GE L, BREWSTER C A. Informational institutions in the agrifood sector: meta-information and meta-governance of environmental sustainability[J]. Current Opinion in Environmental Sustainability, 2016, 18: 73-81.

[13] RUAN T, DONG X, WANG H, et al. Evaluating and comparing web-scale extracted knowledge bases in Chinese and English[C]//Joint International Semantic Technology Conference, November 11-13, 2015, Yichang, China. Heidelberg: Springer, 2015: 167-184.

[14] FAN W F, GEERTS F. Foundations of data quality management[M]. San Francisco Bay: Morgan and Claypool Publishers, 2012: 1-217.

[15] SOMA K, TERMEER C J A M, OPDAM P. Informational governance-a systematic literature review of governance for sustainability in the information age[J]. Environmental Science & Policy, 2016, 56: 89-99.

[16] RUAN T, LI Y, WANG H, et al. From queriability to informativity, assessing

"quality in use" of DBpedia and YAGO[C]//International Semantic Web Conference, October 17-21, 2016, Kobe, Japan. Heidelberg: Springer, 2016: 52-68.

[17] FLECKENSTEIN M, FELLOWS L. Modern data stratery[M]. Heidelberg: Springer, 2018: 63-76.

作者简介

阮彤(1973-),女,博士,华东理工大学计算机技术研究所教授、所长,主要研究方向为数据治理、自然语言处理、知识图谱等。

张知行(1996-),男,华东理工大学计算机科学与技术系硕士生,主要研究方向为信息提取、自然语言处理和知识图谱。

邱家辉(1995-),男,华东理工大学计算机科学与技术系硕士生,主要研究方向为自然语言处理、知识图谱等。

叶琪(1976-),女,博士,华东理工大学计算机科学与技术系讲师,主要研究方向为数据治理、知识图谱等。

收稿日期: 2018-11-02

基金项目: 国家自然科学基金项目(No.61772201)

Foundation Item: The National Natural Science Foundation of China(No.61772201)