

医疗数据治理关键技术综述

摘要

高质量、多模态的数据资源是驱动医疗人工智能模型研发与临床应用的基础。然而，医疗数据具有多源异构、稀疏缺失、隐私敏感和标注不足等特征，直接影响模型的泛化能力、可解释性和临床适用性。医疗数据治理旨在围绕数据全生命周期，提升数据的互操作性、可用性、质量可控性和合规流通能力。本文以医疗数据的全生命周期为主线，系统综述了关键治理技术的演进路径。首先，探讨了在数据采集环节，多模态协同获取与生成式合成如何缓解样本不足问题；其次，聚焦数据标准化与质量控制，剖析了医疗数据在语义转换、智能清洗、自动化质量评估及人机协同标注等方面的主要进展；再次，针对跨机构数据共享与利用需求，讨论了隐私计算、联邦学习和区块链等技术在合规治理中的应用；最后，结合大模型应用背景，归纳了当前医疗数据治理在数据偏见、模型幻觉、跨模态对齐等方面面临的现实挑战，并对数据治理的发展趋势进行了展望。

关键词

医疗数据治理；数据质量；数据标准化；隐私计算

中图分类号：

A Review of Key Technologies for Medical Data Governance

Abstract

High-quality, multimodal data resources serve as the cornerstone for advancing the research and development of medical artificial intelligence models and their clinical translation. However, medical data is characterized by multi-source heterogeneity, sparsity and missing values, privacy sensitivity, and inadequate annotation, all of which directly undermine the generalization capability, interpretability, and clinical applicability of the models. Medical data governance aims to enhance data interoperability, usability, quality controllability, and compliant circulation capacity across the entire data lifecycle. Taking the full lifecycle of medical data as the core thread, this paper systematically reviews the evolutionary trajectory of key governance technologies. First, it examines how multimodal collaborative acquisition and generative data synthesis address the challenge of sample scarcity in the data collection stage. Second, focusing on data standardization and quality improvement, it analyzes the major advances in medical data processing, including semantic transformation, intelligent data cleaning, automated quality assessment, and human-machine collaborative annotation. Third, in response to the demand for cross-institutional data sharing and utilization, it discusses the applications of technologies such as privacy computing, federated learning, and blockchain in compliance-focused governance. Finally, against the backdrop of large-scale model applications, this paper summarizes the practical challenges facing current medical data

governance in terms of data bias, model hallucination, and cross-modal alignment, and provides an outlook on the future development trends of data governance.

Key words

medical data governance, data quality, data standardization, privacy-preserving computation

0 引言

随着数字健康技术的发展,人工智能(Artificial Intelligence, AI)已逐步应用于疾病辅助诊断、预后评估、药物研发和医院管理等场景,上述应用对医疗数据的规模、质量、标准化程度和合规利用能力提出了更高要求。伴随中央政治局关于“推动人工智能健康有序发展”的战略部署^[1],以及国家数据局《高质量数据集建设指引》的发布^[2],数据已被明确为关键生产要素。尽管医疗机构积累了大量病历文本、医学影像、检验检查和时序数据,但数据孤岛、互操作性不足、质量不稳定和隐私保护压力仍是限制医疗AI应用的重要因素。

在此背景下,医疗数据治理的重要性日益凸显。既有研究已从医疗大数据智能分析需求出发,对医疗数据治理的目标、对象和关键环节进行了较早探索。医疗数据治理的核心在于构建面向智能分析的高质量数据基础,需要围绕数据标准、数据质量、主数据管理、元数据管理和数据安全等方面形成系统化治理能力,为医疗大数据分析和人工智能应用提供可信数据支撑^[3]。因此,医疗数据治理并非单一的数据清洗过程,而是围绕医疗数据的采集、接入、标准化、清洗、评估、标注、流通、共享和安全使用等全过程,建立一套能够支撑数据可信供给与合规利用的技术体系和治理机制。其根本目标,在于提升数据

的可发现性、互操作性、完整性、准确性、安全性和任务适配性,使原始、分散、复杂的医疗数据逐步转化为可支撑智能应用和价值释放的高质量数据资源。

从技术演进来看,医疗数据治理已不再局限于传统依赖人工规则和专家经验的处理模式。一方面,深度学习、主动学习和生成式模型的引入,改变了数据清洗、数据增强和标注协同等环节的实现方式;另一方面,大语言模型、隐私计算和联邦学习等技术的发展,也正在推动医疗数据治理从静态处理走向动态优化,从单点工具走向平台化与系统化协同。

尽管近年来国内外围绕医疗数据标准化、数据质量评估、联邦学习和隐私保护等方向已形成一定研究积累,但仍存在两个较为明显的不足:其一,多从单一技术环节展开讨论,缺少面向数据全生命周期的整体性梳理;其二,对大模型背景下医疗数据治理范式变化的讨论仍不充分,尤其在合成数据真实性、自动化质量评估与安全合规流转的内在联系方面关注有限。本文相对于已有医疗数据治理综述的增量贡献主要是:(1)构建了面向全生命周期的医疗数据治理技术演进路径框架;(2)分析了大模型时代合成数据真实性、自动化质量评估与合规流转的内在关联;(3)提出了从规则驱动向人机协同、持续反馈演进的范式转变;(4)补充了监管环境下的合规治理讨论。

本文以医疗数据全生命周期为主线,按照“数据获取/预处理/标注/合规流转与安全治理”的逻辑链条,如图1所示,对

医疗数据治理的关键技术进行系统综述。第1节探讨多源医疗数据采集、异构接入与生成式增强技术；第2节围绕数据质量提升，分析标准化转换、数据清洗、质量评估与人机协同标注等关键环节；第3节

梳理医疗数据在共享流通过程中的安全防护、隐私保护与合规治理机制；第4节讨论当前医疗数据治理面临的主要挑战与发展趋势；第5节做全文总结。

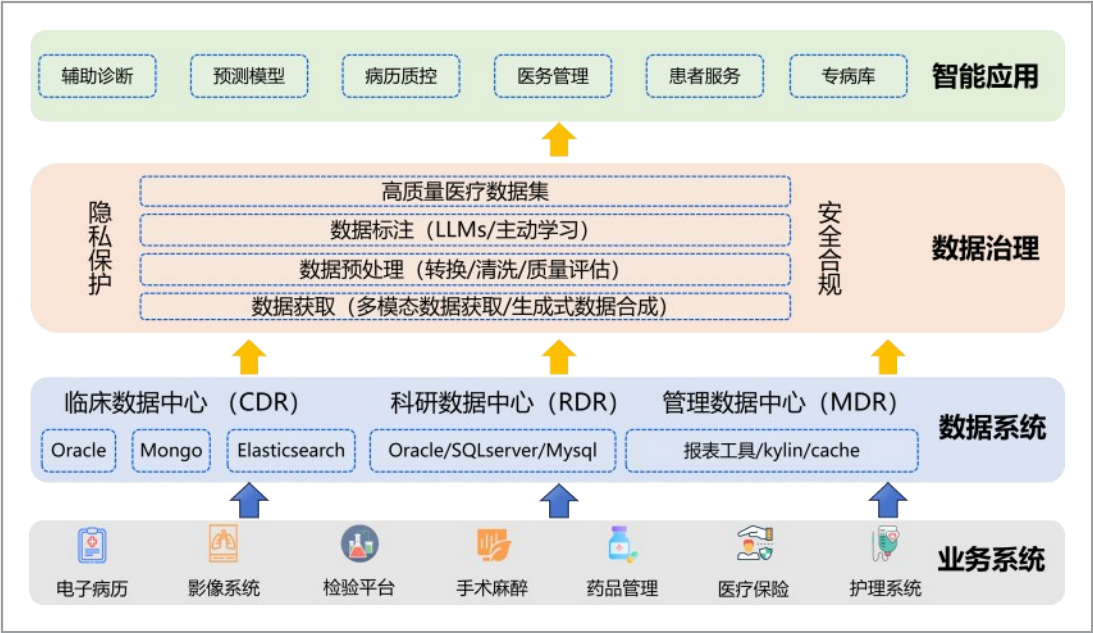


图1 医疗数据全生命周期治理框架

1 医疗数据采集与数据增强

数据采集获取是整个数据治理生态的起点与先决条件，也是后续标准化处理、质量控制与合规流通的前提。与一般行业数据相比，医疗数据来源更加分散、类型更加复杂，既包括电子病历、检验检查、医保结算等结构化或半结构化信息，也包括医学影像、生理信号波形、病理切片以及临床文本等非结构化数据。由于这些数据分别产生于医院信息系统、专科设备、监护终端和区域卫生平台，不同系统在接

口协议、采样频率、记录粒度和编码标准上普遍存在差异，因此，数据采集本身就包含显著的治理需求，而并非单纯的数据汇聚过程。

从临床实际应用看，医疗人工智能对数据的要求不仅体现在规模上，更体现在多模态覆盖度、时序连续性和罕见病例代表性等方面。然而，真实世界医疗数据的获取通常受到伦理审查、隐私约束、临床流程、采集成本以及机构壁垒等多重限制，单纯依赖传统采集方式往往难以满足模型训练和验证的需要。尤其在罕见病、低发事件和复杂并发症等场景中，样本稀缺几乎是普遍现象。基于此，当前医疗数据获

取的技术路径正在由“真实世界直接采集”逐步扩展为“多源协同采集与合成数据增强并行”的双轨模式。前者强调尽可能完整、准确地保留患者真实诊疗过程，后者则致力于在控制风险的前提下补充样本多样性和长尾分布覆盖，从而为后续治理和建模提供更坚实的数据基础。

1.1 多模态医疗数据同步获取与统一接入

单一模态数据难以全景刻画患者的病理生理轨迹。例如，在重症监护、围术期管理和慢病连续监测等场景中，临床决策通常需要综合分析生命体征波形、检验指标、医学影像和病历文本等多源信息。因此，多模态数据采集的关键不只是“采到更多数据”，更在于实现跨设备、跨系统和跨模态的数据协同组织。

在数据获取环节，对齐是基础问题。需要说明的是，多模态医疗数据的“同步”并不局限于狭义的时间戳一致，而是包含至少三个层面的协同：其一是时间维度对齐，即不同设备采集的时序数据在统一时间基准下完成对应，例如同一心动周期内心电信号与有创血压波形的同步；其二是空间维度对齐，主要体现为不同成像设备生成的影像在解剖位置上的配准，例如同一病灶的CT与MRI影像对应；其三是语义维度对齐，即病历文本、检验结果和影像所指向的临床事件、疾病实体或诊疗行为之间能够建立稳定映射。只有在这三个层面形成较好的对应关系，多模态数据才具备后续融合分析和可信建模的基础。

围绕上述目标，近年来多源协同采集技术不断发展。面向高频监测数据，新型物联网医疗中间件可借助边缘网关在数据源头完成初步时间同步、协议转换和标准化封装^[4]，从而减少异构设备直接接入带

来的时序漂移和数据丢失问题。对于大型医院或区域医疗平台，动态数据湖接入技术能够在保留原始信息的基础上，对异构HIS（Hospital Information System）、LIS（Laboratory Information System）和PACS（Picture Archiving and Communication System）中的底层数据进行抽取、汇聚和统一管理^[5]，提高原始数据采集的完备性和一致性。应当指出的是，这类技术虽然提升了接入效率，但其应用效果仍高度依赖接口标准、元数据治理能力以及机构间协同程度，特别是在跨院数据整合场景中，系统异构和语义不一致仍是现实障碍。

1.2 生成式技术驱动的数据合成增强

在真实世界数据难以充分获取、或样本分布呈现明显长尾特征时，合成数据增强正成为医疗数据获取的重要补充路径。与传统意义上的数据增强不同，这里的“合成”不再局限于图像旋转、裁剪、加噪等浅层变换，而是试图利用生成模型学习数据分布特征，生成具有一定真实性和可用性的虚拟样本。其核心目标不是替代真实数据，而是在特定场景下缓解样本不足、类别失衡和隐私受限等问题。

在医学影像领域，改进的生成对抗网络（Generative Adversarial Networks, GANs）以及扩散模型已被用于生成具有特定病灶特征的影像样本，例如罕见肺部疾病X射线图像的合成^[6]。合成增强有助于提升小样本分类或分割任务的训练效果，DeepSynthBody框架^[7]尝试将合成方法应用于心电图、胃肠镜等样本相对稀缺的数据类型，以扩展训练样本的覆盖范围。对于临床文本场景，垂直领域大语言模型可用于生成结构化病历、病例摘要或任务导

向型问答数据^[8]，从而在一定程度上缓解高质量标注文本不足的问题，并降低直接共享真实患者记录所带来的隐私风险。

不过，合成数据增强在医疗场景中的应用仍需保持审慎。一方面，生成样本即便在统计分布上接近真实数据，也未必能够完整反映复杂疾病演变过程和临床决策逻辑；另一方面，模型幻觉、模式崩溃以及长尾样本覆盖不足等问题，仍可能导致合成数据在关键细节上偏离医学事实。特别是在涉及罕见病和高风险诊疗事件时，若缺少临床知识约束和严格验证，合成数据反而可能放大偏差。因此，现阶段更稳妥的定位是将合成数据作为真实世界数据的补充，而非替代真实临床数据，如何建立针对合成数据真实性、可用性和安全性的评价机制，也将成为后续医疗数据治理研究的重要方向^[9]。

2 医疗数据标准化与质量控制

完成多源数据采集后，原始医疗数据通常仍难以直接用于人工智能建模和临床研究。一方面，不同业务系统之间在数据结构、字段定义、编码体系和记录粒度上存在显著差异；另一方面，临床数据在生成过程中又不可避免地受到人工录入习惯、设备状态、采集时点和业务流程差异的影响，因而常伴随缺失、冲突、异常和冗余等问题。对于医疗场景而言，这些问题并非单纯的数据处理瑕疵，而会进一步影响下游模型的稳健性、可解释性与临床可用性。因此，数据治理的重点不应仅停留在“汇聚数据”，而应转向围绕数据可用性、一致性和可信性，尤其在语义层面开展治理。

从技术链路来看，医疗数据在进入分

析和建模环节之前，通常需要经过标准化转换、数据清洗、质量评估以及数据标注等多个步骤。它们并不是彼此割裂的独立模块，而是共同构成了一个持续迭代的质量提升过程：标准化转换解决的是“能否互通”的问题，数据清洗回应的是“是否可靠”的问题，质量评估回答的是“是否可用、可复用”的问题，而标注则直接关系到监督学习任务中的数据表达质量。近年来，随着深度学习和大语言模型在医疗场景中的应用扩展，上述环节的实现方式正在由以人工规则和静态脚本为主，逐步转向规则约束与模型驱动相结合的混合范式。

2.1 语义驱动的医疗数据转换与标准化

异构系统之间的数据互操作，是医疗数据治理中的基础性问题。只有在编码体系、字段语义和数据结构层面建立较为稳定的映射关系，不同机构、不同业务系统产生的数据才有可能进入统一的分析框架。对于医院场景而言，这一问题尤为突出：同一临床概念往往会在 HIS、LIS、PACS 及专科系统中以不同字段形式、不同编码规则甚至不同自然语言表达出现，若缺乏有效的标准化转换机制，后续的数据融合、统计分析和模型训练都难以可靠开展。

在这种情况下，动态 ETL（Extract-Transform-Load）技术^[10]及严格遵循 FAIR（Findable、Accessible、Interoperable、Reusable）原则的语义转换框架^[11-12]，已成为当前医疗数据标准化处理中较为常见的技术路径。与传统静态流程相比，动态 ETL 在规则维护、流程调整和多源适配方面具有更好的灵活性，能够高效地完成字段映射、格式转换和结构归一等基础工作；而 FAIR 导向的语义转

换框架,则更强调数据在可发现性、可访问性、互操作性和可重用性上的整体一致性,适用于跨系统、跨机构数据整合需求较强的场景。

不过,从当前实际来看,标准化转换的难点并不只在结构层面,更集中于语义层面。术语标准化是语义治理的基础环节,主要依赖于国际通用的医学术语系统如 SNOMED CT、LOINC、ICD-10/11 等。然而,直接引入国际标准往往面临本土化适配难题。因此,构建基于本体的临床实践指南术语体系成为研究热点。最新研究通过系统梳理 UMLS、SNOMED-CT、MeSH 及中医临床术语系统,提取并规范指南相关概念,实现了临床术语的标准化表达与逻辑推理^[13]。特别地是,病程记录、手术记录、会诊记录等非结构化临床文本,其内容具有自由书写、多义缩写和上下文依赖强等特点,传统依赖模板和硬编码规则的 ETL 工具往往难以准确完成语义解析。近年来,基于混合专家模型 (Mixture-of-Experts, MoE) 的大模型开始被引入这一环节,例如 Unicorn 等模型^[14]尝试借助多任务学习和参数共享机制,提高异构医疗术语之间的语义映射效率,并改善复杂文本的术语归一化效果。这类方法为医疗数据转换带来了更高的自动化潜力,但其输出结果仍可能受到训练语料质量、领域适配程度以及提示词策略的影响。一旦语义匹配出现偏差,错误将被进一步传递至清洗、评估和建模环节。因此,在大模型参与数据转换时,仍有必要引入标准术语库、知识图谱或人工复核机制,对关键字段和高风险临床概念进行约束与校验。

2.2 基于深度学习与上下文语义校验的数据清洗

在完成初步标准化转换后,数据清洗成为提升医疗数据可用性的关键步骤。需要指出的是,医疗数据中的“脏数据”并不总是表现为一般意义上的格式错误或空值异常,很多问题具有显著的临床语境特征。例如,一项检验结果在统计意义上可能是离群值,但在 ICU 患者中却完全可能对应真实病情;一段看似缺失的主诉数据,也可能与不明原因发热的诊疗路径有关。正因为如此,医疗数据清洗很难完全依赖简单阈值过滤或固定规则,而需要结合特定时序、上下文关系及临床先验进行判断。

传统的数据清洗方法主要依赖统计插补、经验规则和约束检测等技术,在早期结构化数据处理中发挥了重要作用。但随着电子病历、监护时序和多模态临床数据规模的扩大,这些方法在适应复杂分布和动态场景方面的局限逐渐显现。近年来,面向医疗场景的数据清洗开始更多引入深度学习方法,以提升对非线性依赖、隐含异常模式和复杂上下文的建模能力。

针对缺失值问题, LIFE 框架^[15]通过捕捉电子病历中的非线性时间依赖关系,提高了时序数据插补的准确性,尤其适用于存在明显时间相关性的检验数据和病程信息。在异常值检测方面,基于小样本学习的 HoloDetect 模型^[16]、融合统计约束的 SCODED 模型^[17]以及全自动数据清洗框架 Big-MDC^[18]极大地降低了误报率与人工核验成本,相较于仅依赖单点统计特征的传统方法,这类模型更关注异常值在整体上下文中的相对合理性,因此在复杂数据分布下通常具有更好的适应性。

非结构化的病历文书常出现的缩略语不统一、术语误写、逻辑矛盾和时间线紊乱等问题,虽然不一定表现为显式错误,但会直接影响后续的信息抽取和知识建模。大语言模型凭借较强的上下文理解能力,

在文本逻辑校验、医学术语规范化和局部错误修正等方面已显示出一定应用潜力^[19]。不过，对于病历文书清洗而言，模型“看起来合理”的输出并不等同于真实可靠的医学结论。特别是在涉及药品配伍禁忌、诊断路径改变和关键时间节点的场景中，若缺乏标准知识库约束或人工核验，大模型也可能引入新的语义偏差^[20]。因此，现阶段更可行的路径，仍是将其作为规则引擎和人工审核的辅助工具，而不是完全替代临床校验流程。

2.3 医疗数据质量评估

与常见的数据预处理流程不同，医疗数据治理并不是单向推进的线性过程，而更接近一个“处理/评估/反馈/再优化”的闭环。质量评估在这一闭环中具有承上启下的作用：一方面，它用于检验标准化转换和数据清洗是否真正改善了数据质量；另一方面，它也为后续标注、建模、共享和资产化流通提供依据。因此，若缺少明确、可量化的质量评估机制，数据治理就很容易停留在经验层面，难以形成可复用的工程能力。

医疗数据质量评估通常围绕准确性、完整性、一致性和时效性等维度展开^[21]。其中，准确性主要关注数据记录是否真实反映临床事实，例如诊断编码是否符合ICD标准、检验结果单位是否正确；完整性关注关键字段或核心变量是否存在缺失，如确诊糖尿病患者是否缺失血糖相关指标，前列腺癌病例是否缺少PSA数据；一致性则主要考察同一患者在不同系统、不同表单或不同时点上的记录是否存在冲突，例如麻醉记录单与护理记录单中的生命体征是否相互矛盾；时效性则反映数据更新是否满足业务要求，尤其在急诊、重症和实

时预警场景中更为关键。

长期以来，医疗数据质量检查主要依赖人工抽检和规则巡检，这种方式虽然在小规模项目中仍然有效，但在多中心、多模态和持续更新的数据环境中，已经很难满足效率和稳定性的要求。近年来，自动化质量评估框架逐步成为该领域的重要发展方向^[22]。例如，利用知识图谱技术构建医学逻辑校验规则库，能够实现对海量电子病历的自动化关联审查^[23]；基于贝叶斯网络和深度异常检测的评估算法，可对数据清洗后的多模态数据集进行量化打分，生成可视化质量看板^[24]；Medical Data Pecking（MDPT）方法借鉴了软件工程中的单元测试和覆盖率概念，能够自动生成单元测试，从而提高数据质量评估的效率和可靠性^[25]。

需要强调的是，医疗数据质量评估不应仅停留在“是否符合格式规范”这一层面。对于AI驱动的应用场景而言，数据是否适合特定任务、是否会引入系统性偏差，同样是质量评估的重要组成部分。也就是说，质量高并不等于在所有任务中都“可用”，质量评估必须与具体应用目标结合起来理解。例如，对于疾病预测任务，类别分布、标签可靠性和关键特征覆盖度都应纳入评价；而对于多中心研究，机构间记录差异和标准执行偏差则可能比单条记录的局部错误更值得关注。基于这一认识，持续性的质量监测机制^[26]往往比一次性的“验收式评估”更具治理价值。它不仅能够反映数据当前状态，也能帮助发现问题产生的环节，从而为后续优化提供反馈依据。

从治理链条看，质量评估并不是终点。对于大量依赖监督学习的医疗AI任务而言，标注质量本身也是数据质量的重要组成部分。因此，在基础数据质量得到初步保障之后，如何以更高效率、更高一致性

完成专业标注，便成为下一阶段需要解决的问题。

2.4 大语言模型驱动的人机协同数据标注

高质量标注数据是构建医疗人工智能模型的重要前提，但也是最耗费专业资源的环节。与通用场景相比，医疗数据标注具有明显更高的门槛：一方面，标注任务往往需要医学背景支持，例如病灶分割、疾病分型、治疗方案识别和临床关系抽取等均依赖专业判断；另一方面，不同专家之间对复杂病例的理解可能并不完全一致，这使得标注质量控制本身也成为一个问题。因此，如何在保证标注可靠性的同时降低成本、提高效率，已成为医疗数据治理中的关键。

近年来，主动学习为医疗数据标注提供了新的思路。以 HAL^[27]和 HAL-IA^[28]等框架为代表的研究表明，通过将不确定性采样与代表性采样结合，可在较低标注比例下实现较快收敛。这种方法对于影像、时序信号和结构化病例等多类数据均具有一定适用性，尤其适合早期数据集构建阶段的样本筛选。

在主动学习基础上，LLMs 的引入进一步推动了医疗数据标注模式的变化。与传统“人工从零开始标注”不同，越来越多的研究开始尝试采用“模型预标注—专家复核”的协同路径。以 MILO 等方法^[29]为例，LLMs 可在标注前提供候选标签、抽取初步实体关系，或在标注过程中充当交互式辅助工具，从而减少重复性劳动并提高处理效率。对于多模态场景，这类方法的价值更为突出，因为它能够在文本、

影像描述和结构化记录之间建立初步联系，帮助标注员快速定位重点区域或关键事件。

除模型层面的改进外，面向实际工程场景的标注过程管理也在不断发展。针对众包标注噪声问题，基于标注者行为相似度的噪声校正算法^[30]可用于识别低质量标注并提升整体一致性；而 EXACT 等支持版本控制和协同管理的医学影像标注平台^[31]，则为多专家协作、过程追踪和结果复核提供了较为成熟的工程支撑。这些工作说明，医疗数据标注不仅是算法问题，也涉及流程设计、角色分工和质量控制机制。

需要引起重点关注的是，标注人员往往会过度依赖 LLMs 的预设边界，导致人类专家的审核流于形式，进而将大模型的长尾错误（如对微小病灶的漏标）转化为具有误导性的伪金标准。因此，建立基于多专家盲审的冲突仲裁机制，并在模型预标环节强制引入“不确定度量化”以高亮疑似幻觉区域，是保障大模型时代医疗标注质量不可逾越的底线。

2.5 医疗数据治理关键技术比较

为更直观地比较不同技术在医疗数据治理中的适用场景与实施特征，表 1 从治理环节、适用数据模态、主要优势、潜在局限以及资源开销等维度，对当前具有代表性的关键技术进行了归纳。需要说明的是，不同技术并非相互替代关系，而是在真实治理流程中分别作用于数据采集、标准化转换、清洗、评估和标注等不同环节。其实际效果通常还受到业务场景、数据分布、治理目标和机构基础设施条件的共同影响。

表1 典型医疗数据治理技术比较

治理环节	代表性技术/框架	适用数据模态	主要优势	局限性与潜在风险	算力/通信成本
数据采集与增强	新型物联网医疗中间件 ^[4]	时序数据、多模态监测数据	支持多源传感器接入、时间同步和源头标准化封装,有助于提升高频监测数据采集的一致性	依赖设备接口兼容性与时钟同步机制;高频采样场景下对边缘节点存储、计算和网络传输提出较高要求	算力中等,通信较高
	动态数据湖接入技术 ^[5]	异构业务数据、结构化与半结构化数据	支持对 HIS、LIS、PACS 等异构系统数据的抽取与统一归集,便于构建跨系统原始数据底座	元数据治理复杂;跨系统语义映射和权限管理难度较大;若标准控制不足,可能引入后续转换偏差	算力中等,通信中等
	改进 GANs/扩散模型 ^[6-7]	医学影像数据、部分时序数据	可用于补充稀缺病例样本,改善类别失衡,并在一定程度上缓解隐私共享限制	可能存在模式崩溃、长尾样本覆盖不足和病理机制表达不充分等问题;合成样本质量不稳定时影响模型泛化	算力较高,通信较低
	垂直领域 LLMs 合成 ^[8]	临床文本、结构化病历数据	便于生成任务导向型文本样本,提升语料多样性,对隐私敏感文本场景具有补充价值	可能出现医学事实错误、逻辑不一致和隐性偏见;易将错误知识引入训练语料	算力很高,通信较低
	Dynamic-ETL ^[10]	异构结构化/半结构化数据	自动化完成部分 ETL 流程,规则可读性较好,便于部署和维护	仍依赖预设规则和一定编码能力;面对复杂文本语义时适应性有限;存在规则迁移成本	算力较低,通信较低
数据转换与标准化	FAIR 导向语义转换框架 ^[11-12]	多源异构数据	有助于提升数据可发现性、互操作性和重用性,适用于标准化数据转换	标准映射和本体构建成本较高;跨机构实施时需较强治理协同;短期落地复杂度较大	算力中等,通信中等
数据清洗	MoE/LLM 驱动语义映射 ^[14]	临床文本、异构医疗术语数据	对复杂语义匹配和术语归一化具有较强适应性,可提升非结构化文本转换效率	存在语义漂移、错误映射和可解释性不足问题;在高风险场景中仍需人工复核和知识库约束	算力很高,通信较低
	LIFE 框架 ^[15]	电子病历时序数据、检验数据	能够建模非线性时间依赖关系,提升缺失值插补效果	模型依赖训练数据质量;极端临床状态下的稳健性仍需验证;结果解释性相对有限	算力中等,通信较低
	HoloDetect 模型 ^[16]	小样本数据、监测数据	适用于异常值检测,可在有限样本下降低误报率	可能难以区分真实极端生理值与设备噪声;跨机构迁移能力有待进一步评估	算力中等,通信较低
	SCODED 模型 ^[17]	结构化数据	可较好识别系统性错误,尤其适用于缺失和排序异常检测	对单字段细粒度错误识别能力有限;若清洗阈值设置不当,可能误伤少见但真实的临床记录	算力较低,通信较低

续表

治理环节	代表性技术/框架	适用数据模态	主要优势	局限性与潜在风险	算力/通信成本
质量评估	Big-MDC 框架 ^[18]	结构化数据	可从数据中自动挖掘上下文规则,减少对人工规则库的依赖	自动修正规则可能覆盖罕见病例特征;在临床长尾分布场景中需谨慎设置修正策略	算力较低,通信较低
	Metric Hub 框架 ^[24]	多模态数据	可根据任务和场景匹配质量评估指标,有助于区分数据缺陷与模型问题	过度依赖统一评分,可能忽略具体任务对数据适用性的差异化要求	算力中等,通信较低
	MDPT 方法 ^[25]	结构化医疗数据	借鉴单元测试思想,支持自动化质量检测,并可结合外部知识进行上下文敏感测试	依赖测试规则设计质量;在异质性较强数据场景中可能需要大量本地化调整	算力较高,通信较低
	ODM-DQA-Reporter ^[26]	ODM 格式临床研究数据	可自动生成数据质量报告,适合完整性和句法正确性等基础质量检查	评价维度相对有限;对于复杂临床语义错误的识别能力不足;持续监测能力仍需增强	算力较低,通信较低
数据标注	HAL/HAL-IA ^[27, 28]	多模态数据	通过主动学习优先选择高价值样本,提升有限标注资源下的标注效率	可能引入样本选择偏差;若预标注结果缺乏严格复核,易放大模型在长尾样本上的误差	算力较高,通信较低
	MILO ^[29]	多模态数据	支持 LLM 预标注+人工复核协同模式,可提高复杂任务标注效率	存在自动化偏见风险;标注员过度依赖模型建议,影响标注独立性	算力很高,通信较低
	WSNC 类噪声校正方法 ^[30]	众包结构化标注数据	有助于降低众包环境中的标注噪声,提高结果一致性	标注者相似性估计可能受专业背景和偏见影响,对复杂或罕见病例的适用性有限	算力中等,通信较低
	EXACT 等标注平台 ^[31]	医学影像、多模态标注数据	支持版本控制、协同标注与过程追踪,便于质量管理和结果复核	平台本身不直接解决医学判断分歧;对于高复杂度任务仍依赖资深专家介入	算力中等,通信中等

算力/通信成本为相对定性评价,分为较低、中等、较高和很高。

3 医疗数据共享与合规治理

贯穿医疗数据全生命周期的安全与隐私保护,是医疗数据治理不可回避的基础问题。与一般行业数据相比,医疗数据不仅具有高度个体敏感性,还直接关联患者

身份、疾病状态、诊疗过程和用药信息,一旦在采集、存储、交换或使用过程中发生泄露、滥用或误用,可能同时带来个人权益受损、机构合规风险上升以及社会信任受挫等多重后果。因此,医疗数据治理不仅要关注数据质量,而且要在采集、存储、处理、开放和使用等环节建立统一的数据模式、权限控制和安全策略,以降低

数据异构性和使用风险^[32]。

随着《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》的印发，强化数据要素的安全合规流转成为重要趋势。基于匿名化的大数据管理框架ABDM^[33]在保障分析性能的同时，可在一定程度上降低隐私风险，但医疗数据具有高维、强关联和可组合识别等特点，单纯删除直接身份标识并不能完全消除再识别风险。当前研究的重点在于，如何在原始数据不出域、隐私边界可控的前提下，实现跨机构、跨系统的数据协同利用与可信共享。

联邦学习通过在边缘侧（各医院本地）训练模型并仅交互加密梯度参数，为缓解机构间隐私顾虑提供了一种技术路径。但联邦学习并非天然安全，仍可能受到梯度反演、成员推断和模型投毒等攻击，通常需要与安全聚合、差分隐私或可信执行环境结合使用。联邦学习已在跨机构影像分析、电子病历建模和多中心研究中得到初步验证，但仍面临数据异质性和通信开销等工程挑战^[34]。

区块链技术主要用于数据授权、访问日志、操作审计和流转溯源^[35, 36]。由于医疗原始数据体量大、更新频繁，实际应用中更适合采用“链上存证、链下存储”的方式。总体而言，医疗数据安全治理不是单项技术的简单叠加，而是覆盖数据采集、处理、共享和销毁全过程的综合机制，其关键在于形成权限清晰、过程可控、行为可审计、风险可追责的治理框架。

随着人工智能模型训练和评估对数据质量、数据来源和数据合规性的要求不断提高，面向人工智能的数据治理应围绕数据供给、数据质量、数据安全、数据合规、数据标注、数据资产管理和模型应用反馈等环节构建系统框架，以支撑 AI 模型训

练、评测和持续迭代^[37]。该框架对医疗场景具有直接启示：医疗数据治理不仅要解决数据“能否共享”的问题，还要回答数据“是否适合用于医疗 AI 模型训练”“是否可溯源”“是否满足伦理和合规要求”等问题。因此，医疗数据治理需要从传统数据管理进一步扩展到面向模型开发全过程的数据责任管理。

4 医疗数据治理的问题与发展方向

尽管医疗数据治理相关技术不断发展，但在大模型进入医疗文本处理、多模态建模和辅助决策等场景后，数据治理面临的问题更加复杂。当前挑战不仅体现在数据规模和格式统一方面，也涉及数据代表性、生成内容可信性、隐私保护、标准协同和责任界定等问题。

4.1 大模型时代医疗数据治理面临的问题挑战

第一，数据获取受限与合规成本上升。大模型训练和评估依赖高质量医疗语料和多模态数据，但医疗数据受隐私保护、伦理审查和授权许可等因素制约，跨机构共享成本较高。公开医疗数据集在规模、疾病覆盖和时效性方面也难以完全满足垂域模型训练需求^[38]，即便是 Med-PaLM 等代表性医学大模型，其性能仍高度依赖高质量临床语料与专家反馈数据^[39]。

第二，数据偏倚和长尾样本不足。不同地区、不同层级医疗机构在疾病谱、设备条件和记录习惯上存在差异，可能导致训练数据分布不均衡。对于罕见病、复杂并发症和特殊人群，样本不足会影响模型外部泛化能力^[40]。

第三，模型幻觉与合成数据失真。大

语言模型在医学文本生成、辅助标注和知识抽取中具有应用潜力，但也可能生成缺乏依据或与医学事实不一致的内容^[41]。检索增强生成可在一定程度上降低生成错误风险，但其效果依赖知识库质量、检索策略和结果验证机制。合成数据同样需要经过真实性、可用性和隐私风险评估，不能直接替代真实世界数据。

第四，跨模态语义对齐仍不充分。医疗诊疗过程涉及文本、影像、信号、组学和结构化记录等多种数据类型，不同模态在时间粒度、语义表达和采集标准上存在差异。如何实现跨模态数据的一致表示和可靠关联^[42]，仍需进一步研究，跨模态数据治理与对齐是支撑通用医疗大模型的关键基础^[43]。

4.2 迈向自动化与智能体化的数据治理发展方向

未来医疗数据治理将从项目式处理转向持续化治理。首先，应建设面向医疗垂直模型的高质量数据资源体系，提升数据在疾病谱、机构层级、模态类型和人群结构上的覆盖度。其次，应建立动态质量监测机制，对数据缺失、异常、漂移和标签一致性进行持续评估。再次，应推进隐私计算、联邦学习和审计追踪机制的协同应用，在保障隐私安全的同时支持跨机构联合分析。最后，应完善数据质量、标注规范、模型训练数据说明、审计日志和责任划分等配套标准，提高医疗数据治理的可重复性和可监管性。

面向元宇宙医疗、远程协同诊疗和数字孪生患者等新型应用，医疗数据治理还需要从静态数据管理扩展到多场景、动态交互数据治理。已有研究表明，元宇宙环境下的智能医疗模型需要处理多场景异构数据、实时交互数据和复杂环境扰动，因

此对数据采集一致性、跨场景语义对齐、模型鲁棒性和隐私保护提出了更高要求^[44]。

5 结论

医疗数据治理贯穿数据采集、接入、标准化、清洗、评估、标注、共享和应用全过程，是提升医疗人工智能模型可靠性和临床适用性的基础工作。本文从全生命周期视角出发，梳理了医疗数据采集与增强、标准化处理、质量控制、协同标注、隐私保护和合规流转等关键技术，并分析了大模型背景下医疗数据治理面临的主要问题。

总体来看，医疗数据治理正在由依赖人工规则和单点工具的处理方式，逐步转向规则约束、模型辅助和持续反馈相结合的治理模式。生成式模型、大语言模型、隐私计算和联邦学习等技术拓展了数据治理手段，但医疗场景对数据真实性、隐私安全、可解释性和责任可追溯性的要求并未降低。未来，医疗数据治理需要进一步加强高质量多模态数据资源建设、动态质量监测、人机协同审核、隐私保护计算和可信审计机制，并在统一标准和制度规范支撑下，推动医疗数据的安全、合规和可持续利用。

参考文献：

[1] 中央政治局. 习近平在中共中央政治局第二十次集体学习时强调坚持自立自强突出应用导向推动人工智能健康有序发展[EB/OL]. (2025-04-26) [2026-03-27]. <https://www.gov.cn>.
[2] 国家数据局. 高质量数据集建设指引[R]. 北

- 京: 国家数据局, 2025.
- [3] 阮彤, 邱加辉, 张知行, 等. 医疗数据治理——构建高质量医疗大数据智能分析数据基础[J]. 大数据, 2019, 5(1): 2019002. DOI: 10.11959/j.issn.2096-0271.2019002.
- [4] Lv J, Kim B-G, Parameshachari B D, et al. Large Model-driven hyperscale healthcare data fusion analysis in complex multi-Sensors [J]. Information Fusion, 2025, 115: 102780.
- [5] Kalamkar S, Amalanathan G M. MDA-ViT: Multimodal image fusion using dual attention vision Transformer [J]. Multimedia Tools and Applications, 2024, 84 (21): 23701-23723.
- [6] Prakash E, Valanarasu J M J, Chen Z, et al. Evaluating and Improving the Effectiveness of Synthetic Chest X-Rays for Medical Image Analysis [Z]. arXiv, 2024.
- [7] Thambawita V, Hicks S A, Isaksen J, et al. DeepSynthBody: the beginning of the end for data deficiency in medicine[C]. 2021 International Conference on Applied Artificial Intelligence (ICAPAI), 2021: 1-8.
- [8] Vardhan M, Nathani D, Vardhan S, et al. Large language models as synthetic electronic health record data generators [C]. 2024 IEEE Conference on Artificial Intelligence (CAI), 2024: 804-810.
- [9] Gonzales A, Guruswamy G, Smith S R. Synthetic data in health care: A narrative review[Z]//PLOS Digital Health. 2023.
- [10] Ong T C, Kahn M G, Kwan B M, et al. Dynamic-ETL: A hybrid approach for health data extraction, transformation and Loading [J]. BMC Medical Informatics and Decision Making, 2017, 17(1).
- [11] Sinaci AA, Gencturk M, Teoman HA, et al. A Data Transformation Methodology to Create Findable, Accessible, Interoperable, and Reusable Health Data: Software Design, Development, and Evaluation Study [J]. J Med Internet Res. 2023; 25:e42822.
- [12] Pacaci A, Gonul S, Sinaci A A, et al. A Semantic Transformation Methodology for the Secondary Use of Observational Healthcare Data in Postmarketing Safety Studies [J]. Frontiers in Pharmacology, 2018, 9.
- [13] 任慧玲, 李亚子, 李晓瑛, 冀玉静, 邓盼盼. 我国临床医学术语体系建设路径与实践探索[J]. 医学信息学杂志, 2022, 43(8):2-7.
- [14] Tu J, Fan J, Tang N, et al. Unicorn: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration [J]. Proceedings of the ACM on Management of Data, 2023, 1(1): 1-26.
- [15] Heilbroner S P, Carter C, Vidmar D M, et al. LIFE: A Deep Learning Framework for Laboratory Data Imputation in Electronic Health Records [Z]. Cold Spring Harbor Laboratory, 2023.
- [16] Heidari A, McGrath J, Ilyas I F, et al. HoloDetect: Few-Shot Learning for Error Detection [J]. arXiv, 2019.
- [17] Yan J N, Schulte O, Zhang M, et al. SCODED: Statistical Constraint Oriented Data Error Detection [C]. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020: 845-860.
- [18] Nadjeh N, Abdellaoui S, Nader F. Big-MDC: A Multi-Dimensions Cleaning Framework to Improve Big Data Quality [J]. International Journal of Data Science and Analytics, 2025, 20(6): 5869-5893.
- [19] 赵从朴, 朱溥珏, 陈广伟, 张宏伟, 王瑞, 陈政. 基于大语言模型的病历文书智能纠错方法应用与效果研究[J]. 中国数字医学, 2025, 20(11): 96-102.
- [20] YU E, CHU X, ZHANG W, et al. Large Language Models in Medicine: Applications, Challenges, and Future Directions

- [J]. International Journal of Medical Sciences, 2025, 22(11): 2792–2801.
- [21] Schwabe D, Becker K, Seyferth M, et al. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review[J]. npj Digital Medicine, 2024, 7(1). DOI: 10.1038/s41746-024-01196-4.
- [22] Pezoulas V C, Kourou K D, Kalatzis F, et al. Medical data quality assessment: On the development of an automated framework for medical data curation[J]. Computers in Biology and Medicine, 2019, 107: 270–283.
- [23] Fadlallah H, Kilany R, Haber M, et al. Automated big data quality assessment using knowledge graph embeddings[J]. International Journal of Data Mining, Modelling and Management, 2025, 17(4): 383–405.
- [24] Becker K, Oppelt M P, Zech T S, et al. Metric Hub: A metric library and practical selection workflow for use-case-driven data quality assessment in medical AI[A/OL]. arXiv, 2026. <https://arxiv.org/abs/2601.22702>. DOI: 10.48550/ARXIV.2601.22702.
- [25] Girshovitz I, Ambus A, Shahar M, et al. Medical Data Pecking: A Context-Aware Approach for Automated Quality Evaluation of Structured Medical Data[Z]// Machine Learning. 2025.
- [26] Sürer A, Riepenhausen S, Storck M, et al. ODM-DQA-Reporter: A Generic Approach to Assess and Monitor Basic Data Quality of Medical Research Data in Operational Data Model (ODM) Format [M/OL]//Studies in Health Technology and Informatics. IOS Press, 2022. DOI: 10.3233/shti220229.
- [27] Wu X, Chen C, Zhong M, et al. HAL: Hybrid active learning for efficient labeling in medical Domain [J]. Neurocomputing, 2021, 456: 563–572.
- [28] Li X, Xia M, Jiao J, et al. HAL-IA: A Hybrid Active Learning framework using Interactive Annotation for medical image Segmentation [J]. Medical Image Analysis, 2023, 88: 102862.
- [29] Wang Y, Stevens D, Shah P, et al. Model-in-the-Loop (MILO): Accelerating Multimodal AI Data Annotation with LLMs [Z]. arXiv, 2024(2024).
- [30] Hu Y, Jiang L, Zhang W. Worker Similarity-based noise correction for Crowdsourcing [J]. Information Systems, 2024, 121: 102321.
- [31] Marzahl C, Aubreville M, Bertram C A, et al. EXACT: A collaboration toolset for algorithm-aided annotation of images with annotation version Control [J]. Scientific Reports, 2021, 11(1).
- [32] 马朝辉, 裴瑞华, 谭昊翔, 等. 大数据治理的数据模式与安全[J]. 大数据, 2016, 2(3): 2016033. DOI:10.11959/j.issn.2096-0271.2016033.
- [33] Kim J, Jeong J, Kim J, et al. ABDM: Anonymity-Based Big Data Management for Protecting Healthcare Data from Privacy Breach [J]. IEEE Network, 2025, 39(1): 298–305.
- [34] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172–180.
- [35] Prasanna, GL. Privacy Preserving Data Sharing Cloud-Based Healthcare Systems[J]. International Journal of Scientific Research in Engineering and Management, 2025, 09(01): 1–9.
- [36] Borse Y, Sharma D. Blockchain Based e-Healthcare Data Sharing Framework [C]. 2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), 2024: 1–6.
- [37] 李继峰, 张成龙, 刘鑫, 等. 面向人工智能的数据治理框架 [J]. 大数据, 2025, 11(01): 3–20.

DOI: 10.11959/j. issn. 2096-0271.2025004.

[38] Longpre S, Mahari R, Lee A, et al. Con - sent in Crisis: The Rapid Decline of the AI Data Commons[A/OL]. arXiv, 2024. <https://arxiv.org/abs/2407.14933>.

[39] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine[J]. Nature Medicine, 2023, 29(8): 1930-1940. <https://doi.org/10.1038/s41591-023-02448-8>.

[40] Antunes R S, André da Costa C, K ü derle A, et al. Federated Learning for Healthcare: Systematic Review and Architecture Proposal[J]. ACM Transac - tions on Intelligent Systems and Tech - nology, 2022, 13(4): 1-23.

[41] Shaik T, Tao X, Li L, et al. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to Wisdom [J]. Information Fusion, 2024, 102: 102040.

[42] Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medi - cal artificial intelligence[J]. Nature, 2023, 616(7956): 259-265.

[43] Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for elec - tronic health records[J]. npj Digital Medicine, 2023, 6(1). DOI: 10.1038/s41746-023-00879-8.

[44] 朱玖闻, 周玉冰, 斯洪标, 等. 一种高效鲁棒的元宇宙环境下的多场景智能医疗模型研究[J]. 大数据, 2024,10(2):122-139.

录用日期: XXXX-XX-XX

通信作者:

基金项目: 首都卫生发展科研专项(首发2026-2G-40110);国家重点研发计划项目“基于区块链的卫生健康数据可信共享技术及示范应用”(2022YFB2703300)

Foundation Items: Capital's Funds for Health Improvement and Research (No.2026-2G-40110); National Key Research and Development Program of China “Trusted Sharing Technology and Demonstration Application of Health Data Based on Blockchain” (No.2022YFB2703300)

