

医疗领域大模型伦理风险识别、治理及前瞻研究

刘浏^{1,2,3}, 张馨⁴, 张琪琪^{1,2}, 刘洁¹, 刘子裕¹, 王惠玲¹, 易汉希², 王维圆¹,
Ousmane Diabate^{2,5}, 王俊普^{1,2,3*}

(1. 中南大学湘雅医院, 长沙 410008; 2. 中南大学湘雅基础医学院, 长沙 410008; 3. 中南大学商学院, 长沙 410083;
4. 马斯特里赫特大学医学中心, 马斯特里赫特 6211LK; 5. 马里医院, 巴马科 999053)

摘要: 医疗领域正经历以多模态大模型为特征的智能化转型, 然而技术赋能与伦理风险共生并存, 医疗领域大模型的深度应用使伦理风险日益凸显, 亟需聚焦医疗领域大模型, 全面识别其伦理风险并探索有效的治理路径。本文深入探讨了医疗领域大模型应用过程中存在的数据隐私风险、算法决策风险、主体关系风险、社会公平风险等四大核心伦理风险, 并结合典型案例加以解析; 提出了基于“数据-算法-应用-法律”四位一体的医疗领域大模型治理框架, 涵盖构建数据治理体系、创新算法治理机制、建设临床应用规范、完善法律监管框架4个方面; 分析了医疗领域大模型发展面临的关键技术挑战和政策挑战。最后, 展望了医疗领域大模型的未来方向, 包括探索基于区块链医疗数据确权、开发轻量化模型普惠基层医疗、构建“政产学研医”协同生态系统, 以期为推动医疗领域大模型技术规范健康发展、保障患者权益及完善医疗伦理治理体系提供理论与实践支撑。

关键词: 大模型; 医疗大模型; 伦理风险; 风险识别; 治理路径
中图分类号: R-052; TP18 **文献标识码:** A

Ethical Risk Identification, Governance, and Foresight of Medical Foundation Models

Liu Liu^{1,2,3}, Zhang Xin⁴, Zhang Qiqi^{1,2}, Liu Jie¹, Liu Ziyu¹, Wang Huiling¹, Yi Hanxi²,
Wang Weiyuan¹, Ousmane Diabate^{2,5}, Wang Junpu^{1,2,3*}

(1. Xiangya Hospital of Central South University, Changsha 410008, China; 2. Xiangya School of Basic Medical Sciences, Central South University, Changsha 410008, China; 3. Business School, Central South University, Changsha 410083, China;
4. Maastricht University Medical Center, Maastricht 6211LK, Netherlands; 5. Hospital Du Mali, Bamako 999053, Mali)

Abstract: The medical field is undergoing an intelligent transformation characterized by the emergence of multimodal foundation models. While technological empowerment brings unprecedented opportunities, it is accompanied by profound ethical risks. The deep application of medical foundation models has amplified such risks, necessitating comprehensive identification and effective governance strategies. This study systematically examines four core categories of ethical risks in medical foundation model applications: data privacy risks, algorithmic decision-making risks, subject-relation risks, and social equity risks, illustrated through representative cases. To address these challenges, a data-algorithm-application-law integrated governance framework is proposed, encompassing the establishment of robust data governance systems, innovation of algorithmic governance mechanisms, formulation of clinical application

收稿日期: 2025-07-19; 修回日期: 2025-08-26

通讯作者: *王俊普, 中南大学湘雅医院 / 湘雅基础医学院副教授, 研究方向为病理学与数智病理; E-mail: wang-jp2013@csu.edu.cn

资助项目: 芙蓉实验室项目(2024PT5111)

本刊网址: sscae.engineering.org.cn

standards, and improvement in legal and regulatory frameworks. Furthermore, the study analyzes key technological and policy challenges constraining the development of medical foundation models. Looking ahead, the study outlines potential future directions, including the exploration of blockchain-based medical data ownership confirmation, development of lightweight models to promote equitable healthcare at the grassroots level, and construction of collaborative ecosystems integrating government, industry, academia, research, and healthcare. These efforts are intended to provide theoretical foundations and practical pathways for fostering the normative and sound development of medical foundation model technologies, ensuring patient rights, and enhancing the ethical governance system in the healthcare domain.

Keywords: foundation model; medical foundation model; ethical risk; risk identification; governance approaches

一、前言

人工智能 (AI) 正在深刻变革医学^[1-3]。随着生成式 AI 技术的突破性发展^[4], 医疗领域正经历以多模态大模型为特征的智能化转型^[5-7]。以国外谷歌公司 Med-PaLM 医疗大模型^[8]、斯坦福大学 MUSK 病理大模型^[9]、OpenAI 公司 ChatGPT 衍生医疗模型以及国内深圳市腾讯计算机系统有限公司的觅影医疗大模型、科大讯飞股份有限公司的星火医疗大模型、上海交通大学医学院附属瑞金医院与华为技术有限公司联合发布的瑞智病理大模型^[10]、杭州深度求索人工智能基础技术研究有限公司的 DeepSeek 衍生医疗模型等为代表的医疗领域大模型, 在医疗文本处理、医学图像分析、药物与医疗器械研发、医学教育及健康管理等方面展现出颠覆性发展潜力^[11-13]。例如, 谷歌公司的 Med-PaLM 2 模型在医学问答任务中首次通过美国医师执照考试, 准确率超过 90%^[8]。然而, 技术赋能与伦理风险共生并存, 医疗领域大模型带来的伦理风险逐渐凸显。2015 年, 英国深度思考公司 (DeepMind) 未经充分授权获取了 160 万份肾衰竭患者的医疗数据用于 AI 训练, 引发全球首例医疗 AI 数据滥用诉讼; 2019 年, 《Science》期刊披露了美国某医疗中心重症监护筛选算法的系统性歧视, 导致黑人患者需比白人多患病 40% 才能获得同等护理资格^[14]; 2023 年, 美国芝加哥大学与谷歌公司被患者提起集体诉讼, 指控其医疗数据去标识化技术存在漏洞等。诸如上述医疗领域大模型带来的伦理风险, 对患者权益、医疗质量和公平性等构成挑战, 亟待深入研究和有效治理^[15,16]。

医疗领域大模型指基于海量多模态医疗数据 (如电子健康档案 (EHR)、医学影像、基因组数据等) 训练而成、具有高参数规模和强泛化能力的 AI 大模型, 不仅限于医疗大模型、病理大模型等, 能够支持辅助诊断、治疗方案生成、医学问答、健康管理等医疗领域任务。我国高度重视医疗领域大模

型的规范健康发展。2017 年发布的《新一代人工智能发展规划》^[17]首次将智能医疗纳入重点领域, 提出了构建 AI 安全可控治理体系; 2023 年发布的《生成式人工智能服务管理暂行办法》^[18]明确了“包容审慎和分类分级”的监管原则, 并持续开展生成式 AI 服务备案工作, 已备案的模型和算法数量已达 100 多个; 2025 年的《政府工作报告》^[19]中明确指出, 持续推进“AI+”行动, 支持大模型广泛应用, 促进平台经济规范健康发展; 我国提出参考《人工智能医用软件产品分类界定指导原则》^[20], 将大模型辅助诊断纳入三类医疗器械监管范畴等, 这都充分体现了我国对医疗领域大模型从技术驱动向伦理治理并重的战略转向。当前, 全球医疗领域大模型伦理风险治理呈现区域分化特征: 美国通过美国食品药品监督管理局 (FDA) 的 510 (k) 认证强化了算法透明度, 欧盟通过发布的《人工智能法案》确立了“风险分级”的监管模式, 日本则通过《医疗人工智能开发指南》要求企业建立算法影响评估制度, 而我国正积极探索“技术标准+伦理审查”的本土化路径, 在国际经验与国情需求间寻求平衡。医疗领域大模型伦理风险的识别及治理是推动医疗领域大模型技术规范健康发展的必然要求, 是保障患者权益与医疗安全的现实需求, 也是完善我国医疗伦理治理体系的重要实践。

在此背景下, 深入识别医疗领域大模型面临的伦理风险, 并探索切实可行的治理路径, 兼具学术研究与实践应用价值。本文首先系统剖析医疗领域大模型在数据隐私、算法决策、主体关系与社会公平四大维度的伦理风险, 然后创新性提出“数据-算法-应用-法律”四位一体的治理框架, 最后展望探索基于区块链医疗数据确权、开发轻量化模型普惠基层医疗等未来方向, 强调需通过构建“政产学研医”的协同生态系统, 实现技术创新与伦理约束的良性互动, 为医疗行业的智能化转型提供坚实的伦理基础。

二、医疗领域大模型伦理风险识别

随着医疗领域大模型的深度应用，其引发的系统性伦理风险已从技术争议上升为公共治理议题^[10,16]。这些风险根植于数据隐私^[2,21,22]、算法决策^[23-25]、主体关系^[26,27]、社会公平^[28,29]等方面，呈现出技术内生性与社会外延性交织的复杂特征。数据隐私风险不仅源于敏感医疗信息的规模化处理，更因大模型参数对数据特征的强记忆性而催生新型攻击路径；算法决策风险暴露出医疗AI在技术理性与医学人文性之间的深层矛盾，其“技术黑箱”特性对传统医疗伦理框架构成挑战；主体关系风险则指向人机协同模式下医患信任重构与责任分配困境，折射出技术介入对医疗主体性的解构与重塑；社会公平风险进一步揭示技术普惠愿景与现实资源分配失衡间的鸿沟，形成从数据采集到服务落地的结构性闭环。这些风险并非孤立存在^[30]，而是通过算法逻辑与社会结构的耦合作用形成叠加效应，亟需建立多学科协同的治理体系以应对技术伦理的范式变革。

（一）数据隐私风险

1. 数据泄露风险

在医疗领域，大模型的训练高度依赖多模态敏感数据集，其中涵盖基因组信息，如单核苷酸多态性（SNP）位点，能精准反映个体基因特征；EHR详细记录患者的诊疗全程；医学影像的DICOM文件，包含关键的病理影像细节等。数据泄露风险贯穿于数据的全生命周期，不仅在数据存储环节，因存储介质未加密、权限管理失控等因素，易造成数据明文暴露或被越权访问，在模型推理阶段，也面临严峻挑战。攻击者常采用成员推理攻击，利用模型对训练集和非训练集样本响应的细微差异，精准判断特定患者数据是否被用于模型训练，这可能泄露患者的疾病史、遗传倾向等敏感信息。模型反演攻击同样威胁巨大，攻击者借助模型输出，运用优化算法等手段，反向重构输入数据特征，如从病理报告生成模型的输出中，反推出患者原始的病理切片图像，导致患者隐私暴露。一旦这些敏感医疗数据泄露，将对患者的隐私安全、个人权益造成严重损害，甚至引发连锁的社会与伦理问题。

2. 数据滥用风险

在医疗领域，数据滥用风险正日益凸显，医疗数据的“灰色流通”已演变成系统性难题。目前，数据权属界定极为模糊，医院普遍认为诊疗过程中形成的数据属于其管理和使用范围，部分医疗机构在与制药企业或科研机构合作时，将病历中的部分敏感信息（如基因组信息、慢性病记录等）用于药物研发或临床研究，以推动医学进步和产业发展，如某跨国药企就曾借此获得数十亿融资；科技公司则主张用户自愿上传的数据归自身所有，如某健康手环协议中暗藏“用户授权公司无限期使用数据于商业目的”条款。从法律层面看，《中华人民共和国个人信息保护法》规定个人对其信息享有决定权，但现实中患者签署的知情同意书常覆盖未知用途；匿名化的医疗数据在AI交叉比对下也难以发挥有效作用，且维权成本高昂。即便采用联邦学习等先进技术，通过分布式训练以避免原始数据集中传输，但可梯度更新过程依旧存在隐患，可能泄露患者分布特征。大量研究表明，看似已匿名化的医疗数据，如诊断记录，经机器学习模型重新识别的概率高达60%~80%，进一步加剧了数据滥用风险。

3. 典型案例剖析

2015年，DeepMind与英国皇家自由医院合作引发的数据争议成为全球首个医疗AI数据滥用典型案例。DeepMind未经患者明确同意，获取了160万份肾衰竭患者的完整医疗数据（包括血液检测、用药记录及住院史），名义上用于开发急性肾损伤预警系统Streams，但实际数据使用范围远超肾功能监测需求，且涉及商业算法开发。英国信息专员办公室调查发现，患者未充分知晓个人信息将被如此使用，医院也未给予患者拒绝数据被使用的选择权，以“直接医疗照护”为法律依据规避了欧盟的《通用数据保护条例》（GDPR）的“目的限定原则”，而DeepMind则辩称数据已通过哈希加密去标识化。然而，后续研究证明，此类传统匿名化手段在机器学习环境下极易被破解——通过关联模型输出与公开医保数据库，患者重识别率达75%，暴露了隐私增强技术的局限性。该事件直接促使欧盟修订了《医疗数据跨境流动指南》，强制要求AI训练数据需获得“动态同意”，即患者可逐项授权不同用途。

（二）算法决策风险

1. 临床决策偏差

在医疗领域，临床决策偏差已成为AI应用中不容忽视的问题。社会中普遍存在的偏见和歧视，常常会被AI技术不自觉地复制。这是因为AI算法在训练时，存在统计性偏差强化机制，其优化目标天然倾向于数据丰富的多数群体特征分布。以医疗影像诊断为例，若训练数据中白色人种的医学影像占比较大，那么当模型面对有色人种的医学影像时，就可能出现误诊或漏诊，如曾有用于检测皮肤癌的AI技术对有色人种的诊断准确性远低于白色人种。除数据本身的偏见外，开发者偏见也会影响AI决策。开发者在构建算法时，可能因自身认知局限或固有观念，将某些不合理的判断标准融入其中。区域适应性偏差同样显著，不同地区的医疗环境、疾病谱存在差异，若AI模型未充分考虑这些因素，在不同区域应用时就会出现决策失误。例如，在疾病流行特征不同的地区，使用同一套未经优化的AI诊断模型，极有可能给出不恰当的临床决策，进一步加剧医疗保健的不平等，严重影响患者的诊疗效果与健康权益。

2. 可解释性困境

在医疗领域，大模型的可解释性困境正成为制约其广泛应用的关键瓶颈。大模型基于海量数据与复杂神经网络构建，其内部结构如同精密而神秘的“黑箱”，决策机制高度抽象化、参数化，导致输出的诊断结论和治疗方案难以被医生和患者直观理解。例如，当大模型针对罕见病患者输出治疗建议时，无法清晰呈现是基于哪些基因位点突变、影像特征或临床指标做出的判断，致使医生难以据此评估方案的科学性，患者更无法充分了解治疗的依据，这严重削弱了医疗决策的可信度，也与医疗伦理中强调的知情同意原则背道而驰。为此，欧盟发布了《人工智能法案》，将高风险医疗AI纳入监管范畴，要求其遵循GDPR第22条，提供可追溯的决策链，赋予用户“解释权”。然而，当前主流的解释工具如局部可解释模型无关解释方法（LIME）和Shapley加法解释方法（SHAP），在应对多模态医疗数据时仍存在不足，在跨模态特征整合与解释过程中，容易出现解释不连贯、关键信息缺失等问题，无法满足医疗场景对严谨性、专业性的高标准要求，使大模型的可解释性难题依旧悬而未决。

3. 典型案例剖析

2019年，《Science》期刊曝光的美国某大型医疗中心案例^[4]，深刻揭示了医疗AI中的种族偏见危害。该医疗中心用于筛选重症监护患者的算法，以“未来医疗费用”作为预测目标，而非直接反映健康需求，这一设计缺陷结合社会结构性歧视，易酿成严重后果。由于长期存在的就医障碍、保险覆盖不足等问题，训练数据里黑人患者的医疗费用被严重低估，与实际健康需求脱节。算法却将“费用”错误等同于“疾病严重程度”，致使黑人患者需比白人患者多患病40%，才能获得同等护理资格。在算法部署的2亿次决策中，黑人患者的医疗优先级被系统性低估，生命健康权益遭受损害。这一事件引起广泛关注，推动美国监管机构和学术界对医疗AI偏见问题进行深入讨论。随后，美国食品药品监督管理局（FDA）在《人工智能/机器学习医疗软件行动计划》中提出需加强透明度与偏差监测，白宫科技政策办公室（OSTP）在《人工智能权利法案蓝图》中强调避免算法歧视，旨在从制度层面减少医疗AI中的偏见，保障不同种族患者能获得公平的医疗资源分配与服务。

（三）主体关系风险

1. 医患关系异化

在医疗领域，信任构筑起医患关系的稳固桥梁，患者基于对医生专业能力与职业操守的信赖接受诊疗。然而，AI作为新型“主体”介入医患关系后，这种信任基石正面临动摇。当医生在诊疗过程中频繁参考AI大模型的建议，甚至将部分决策权让渡给算法时，患者对医生专业权威性的认知会被削弱，进而引发信任危机。例如，若患者发现医生在诊断时更多依赖AI输出结论，而非自身的经验判断，可能会质疑医生的专业能力，认为其缺乏独立思考与判断的能力。此外，过度依赖大模型还会造成人文关怀的缺失。AI仅能基于数据与算法提供诊疗方案，无法感知患者的情绪与心理需求。当医生的诊疗流程被AI主导，倾听患者诉求、安抚患者情绪的时间大幅减少，医患之间的情感联结将逐渐弱化。长此以往，患者会产生被忽视、不被重视的感觉，加剧医患关系的紧张程度，使原本充满温度的医患互动，异化为冰冷的数据与机器指令交互，严重影响医疗服务的质量与患者的就医体验。

2. 医疗责任模糊

在医疗领域，智能医疗机器人的应用在带来便利的同时，也使医疗责任归属陷入模糊困境。智能医疗机器人的法律主体资格认定是界定其侵权责任的核心，学界普遍认为技术自主性水平决定责任划分。在弱AI阶段，医疗机器人通常被视作医疗器械产品，一旦出现问题，若因算法缺陷，开发者需担责；若是操作不当，医疗机构则要过失负责。但进入强AI场景，当系统具备独立完成临床决策闭环的能力时，传统归责原则将难以适用，需借助“有限电子人格”理论建立新的责任认定体系。当前，大模型辅助诊断引发的三元责任困境尤为突出。深度学习算法的“黑箱”特性，切断了开发者、使用者与数据提供者之间的过错关联，导致责任难以明确。此外，现有法律框架缺乏基于动态风险分配的协同问责机制，无法依据各方对风险的控制能力和收益比例，精准判定责任。一旦出现医疗纠纷，责任认定便陷入僵局，凸显了法律制度在应对智能医疗发展时的滞后性。

3. 典型案例剖析

2023年，美国Cigna Healthcare公司推出的“Px Dx” AI算法系统在自动化处理医疗索赔时暴露了医疗AI应用的深层次伦理问题。该系统通过分析患者历史数据和诊疗记录进行自动审核索赔，但因算法训练数据中基层医疗案例不足（仅占12%），且未充分考虑罕见病、多并发症等复杂情况，导致约23%的合理索赔被错误拒绝。这一事件引发了连锁反应：患者因索赔被拒质疑医生的专业判断，医患沟通时间缩短40%，而医疗机构陷入“双重追责”困境——保险公司往往通过合同条款或内部政策规避责任；患者则可能援引《平价医疗法案》（ACA）及相关消费者保护法规，起诉医疗机构未尽人工复核义务。更严重的是，算法决策过程完全“黑箱化”，连开发工程师也无法解释特定拒赔逻辑，致使美国医疗责任保险费用同比上涨35%。该案例揭示了医疗AI治理的三大盲区：数据代表性缺陷引发的结构性歧视、人权权责划分模糊导致的法律真空以及算法不可解释性对程序正义的破坏，为全球医疗AI伦理治理提供了警示性样本。

（四）社会公平风险

1. “数字鸿沟”加剧

在医疗领域，“数字鸿沟”加剧成为不容忽视

的社会公平风险。经济发达地区凭借充足的资金、先进的技术以及完备的网络设施，能率先引入先进的医疗大模型，并持续投入资源对其进行优化升级。大型医疗机构凭借丰富的临床数据、优秀的医学人才，得以深入挖掘大模型在疾病早期精准诊断、个性化治疗方案定制等方面的潜力。与之形成鲜明反差的是，偏远地区因网络覆盖不佳、带宽有限，致使医疗大模型部署艰难，即便勉强接入，运行也常因网络问题而卡顿，无法稳定提供高效的服务。小型医疗机构与基层医疗单位在部署医疗AI大模型时，不仅面临高昂的技术采购成本和维护费用，还缺乏能熟练运用、维护大模型的专业人才，无法为患者提供与大医院同等水平的智能化医疗服务。这种地域与机构间的差距，进一步加大了医疗资源分配的不平衡，让欠发达地区和基层患者难以享受到医疗领域大模型带来的红利，使“数字鸿沟”持续加深，严重威胁医疗资源的公平分配和医疗服务的普及化。

2. 资源分配偏差

在医疗领域，资源分配偏差问题日益突出。一方面，算法偏好致使资源错配情况频发。部分医疗领域大模型服务设置了较高的使用门槛，无论是患者端的咨询问诊，还是医疗机构接入端运用模型辅助诊断，都需支付高昂的费用。这使得经济条件欠佳的患者、小型医疗机构无力承担，难以享受大模型带来的精准医疗福利，与医疗公平性原则背道而驰。另一方面，数据采集偏移现象显著。基层医疗机构由于缺乏先进的数据采集设备、专业的数据管理人才以及稳定的技术支持，存在数据采集存在数量不足、质量不高的问题，致使在进行大模型训练时，基层医疗数据占比严重不足，模型对基层常见疾病特征、患者群体特点的学习受限，生成的诊疗方案与建议在基层的适用性大打折扣。从数据采集环节的薄弱到实践中难以契合需求，不同地区在AI及大模型应用上的“数字鸿沟”逐渐形成恶性循环，进一步加剧了医疗资源分配的不均，阻碍医疗公平的实现。

3. 典型案例剖析

2024年，阿里巴巴（中国）有限公司推出的“医疗AI多癌早筛公益项目”在浙江丽水正式启动，并赋能基层医疗。该项目由阿里达摩院（湖畔实验室）与当地医疗机构紧密合作，旨在将前沿医

疗 AI 技术创新性地引入卫生健康领域，通过大规模的日常检查实现多癌早筛。此前，阿里达摩院已联合全球顶尖医疗机构，在胰腺癌早筛研究上取得重大突破，相关成果被《自然医学》期刊发表^[31]。在此次公益项目中，依托阿里达摩院医疗 AI 实验室自研的“达医智影”，运用“平扫计算机断层扫描（CT）+AI”的方式，助力胰腺癌、肝癌、食管癌、胃癌、结肠癌等癌种病灶的发现，并辅助医生对骨质疏松、脂肪肝、肺结节等 13 个病种进行全流程诊疗。该项目从胰腺癌和骨质疏松两个病种率先开展早筛工作，后续将逐步接入更多癌症和慢性病的筛查能力。通过此次合作，期望能提升当地数字健康水平，为百姓带来切实的健康福祉，同时节约医疗开支和医保经费，也为医疗资源不均衡地区的多癌早筛提供可推广的范例。

管理暂行办法》^[18]，明确了发展和安全并重、创新与治理结合的原则，提出了算法备案、数据安全评估等要求，对生成式 AI 服务实行包容审慎的分类分级监管。随着 AI 技术的发展，我国已有多款 AI 医疗软件通过国家药品监督管理局的审批、多款医疗大模型完成算法备案，而对于医疗领域大模型的伦理风险治理还在探索阶段^[32,33]，尤其是在数据隐私、算法决策、主体关系、社会公平等方面面临严峻挑战。基于技术特征与国情需求，本文提出基于“数据-算法-应用-法律”四位一体的治理框架（见图 1），融合治理伦理学与风险结构映射理论，强调在医疗 AI 治理中实现责任可溯、风险可控、权益可护的伦理目标，形成数据驱动算法优化、算法支撑应用落地、应用反馈法律完善、法律保障数据安全的闭环逻辑链，构建符合我国医疗场景的伦理治理体系。

三、医疗领域大模型治理路径构建

2023 年 7 月，我国发布《生成式人工智能服务

（一）构建数据治理体系

1. 隐私计算技术应用

在构建数据治理体系中，隐私计算技术的应用

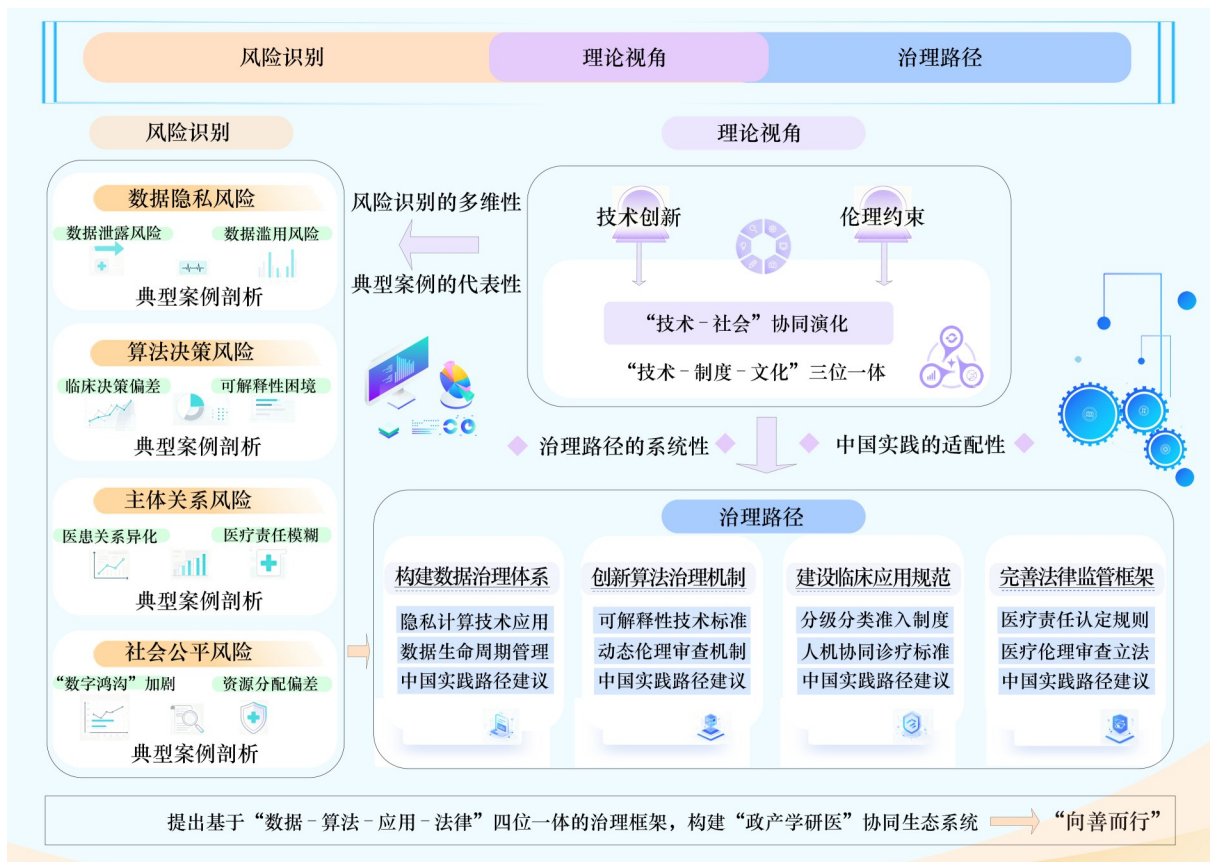


图 1 基于“数据-算法-应用-法律”四位一体的治理框架

是关键环节。首先，采用联邦学习技术，通过分布式训练模式，使各医疗机构在本地数据不共享的前提下共同训练模型，仅交换加密的模型参数更新，确保原始数据始终保留在本地。其次，引入同态加密技术，在数据处理和模型推理过程中对敏感医疗信息（如基因组数据、EHR）进行加密运算，支持在密文状态下完成特征提取和预测分析，避免明文暴露风险。针对数据聚合场景，应用差分隐私技术，在统计查询或模型输出中添加经过数学优化的噪声扰动，确保个体数据无法被反向推断，同时保持整体数据的可用性。此外，结合安全多方计算（MPC）协议，实现跨机构数据联合分析时的隐私保护，通过密码学方法分割计算任务，使各方仅能获得最终结果而无法窥探他方输入。在技术集成层面，设计分层隐私保护架构：底层采用硬件级可信执行环境（TEE）隔离敏感计算过程，中间层部署轻量级边缘计算节点完成数据预处理，应用层则通过动态访问控制与零知识证明技术验证数据使用合法性。最后，建立隐私保护效果量化评估机制，定期测试模型抗攻击能力（如成员推理攻击准确率需低于50%），并通过区块链存证技术记录数据使用全流程，确保隐私计算策略的可审计性。

2. 数据生命周期管理

在医疗数据生命周期管理中，需建立全流程标准化管控机制。在数据采集阶段，采用智能合约技术自动执行数据采集协议，确保患者知情同意书与采集目的严格匹配，并通过元数据标记明确数据来源、采集时间和用途限制。在数据存储阶段，实施“三副本加密存储+区块链存证”策略，原始数据经高级加密标准（AES-256）加密后分布式存储在医疗专有云，同时将数据哈希值上链固化，实现防篡改、可追溯。在数据使用阶段，要部署动态脱敏网关，根据访问者角色自动实施字段级脱敏（如基因数据仅对授权研究人员开放完整序列），并采用属性基加密技术实现细粒度访问控制，确保数据使用严格遵循“最小必要”原则。在数据共享流转阶段，建立数据沙箱环境，使所有外部查询需通过隐私计算中间件处理，仅返回聚合统计结果或模型参数，禁止原始数据导出。在数据销毁阶段，采用物理擦除与逻辑删除双重机制，过期数据经量子随机数覆写7次后自动触发区块链智能合约注销对应访问权限，并在联盟链节点同步删除元数据索引。同

步构建数据质量监测系统，通过预设规则引擎实时校验数据完整性和一致性，异常数据自动触发清洗流程，并通过基于角色的访问控制与基于属性的访问控制的混合模型来实现跨机构数据协作时的精确授权。

3. 我国实践路径建议

在我国医疗数据治理实践中，应构建“国家-区域-机构”三级联动的医疗数据共享安全平台体系。国家级平台可由国家卫生健康委员会牵头建设，基于自主可控的分布式架构，集成隐私计算、区块链和多方安全计算技术，实现跨省份医疗数据的可信流通。区域级节点部署在省级健康医疗大数据中心，负责对接辖区内三级医院和基层医疗机构，通过智能数据网关完成异构数据的标准化处理，并采用联邦学习技术实现区域间的模型协同训练。机构接入层要求医院信息系统改造接口，符合医疗健康数据的共享标准，对接平台时强制启用国产密码算法SM4加密传输和SM9标识认证。“国家-区域-机构”医疗数据共享安全平台的核心功能包括：建立全国统一的医疗数据资源目录，实现数据资产“一档”管理；开发数据沙箱环境，支持科研人员在受控环境下调用脱敏数据；部署AI监管模块，自动检测数据使用合规性（如追踪模型训练时的数据流向）等。在平台运营机制上，实行“数据贡献积分制”，医疗机构按数据质量和数量获得算力资源或科研协作优先权，同时建立数据使用“黑名单”制度，对违规行为实施联合惩戒。在技术保障方面，平台应采用全栈国产化技术路线，并通过等级保护三级认证。配套建设（7×24）h网络安全监测中心，实时防御高级持续性威胁攻击和数据泄漏风险，定期进行渗透测试。

（二）创新算法治理机制

1. 可解释性技术标准

在创新算法治理机制中，可解释性技术标准的实施需要建立多层次的透明化体系。首先，在模型架构层面，采用注意力机制可视化技术，对医学影像诊断模型生成病灶定位热力图，并标注关键特征权重（如CT影像中结节的大小、边缘特征对恶性概率的贡献度）；对于文本诊断模型，则开发决策路径追溯功能，展示从症状输入到诊断结论的推理链条。其次，构建医疗专用的解释性接口，设计双

通道输出模式：面向医生提供专业版解释报告，包含模型决策依据的医学证据链（如参考的临床指南条目、相似病例统计）；面向患者生成通俗版说明，通过可视化图表展示诊断逻辑。在技术实现上，集成 SHAP 值分析、LIME 局部解释等工具，针对不同模态数据（如影像、文本、基因）开发定制化解释模块，确保关键决策因素的可解释覆盖率超过 95%。同时，建立动态公平性监测系统，在模型推理时实时计算不同人群组（按年龄、性别、地域划分）的性能差异指标（如召回率方差不超过 0.05），当检测到潜在偏见时自动触发再训练机制。配套开发解释质量评估工具包，包含临床合理性测试和用户理解度测试，确保解释结果既符合医学规范又具备实际应用价值。最后，将这些技术要求纳入医疗 AI 产品注册审查标准，作为算法备案的强制性内容。

2. 动态伦理审查机制

在动态伦理审查机制建设中，应构建贯穿研发全周期的多维度监管体系。组建跨学科伦理审查委员会，由临床医学专家、AI 伦理学者、法律顾问和数据安全工程师共同构成，采用“双盲评审+交叉验证”机制对模型开发各阶段进行伦理评估。在模型研发阶段，实施“伦理设计前置”原则，要求提交训练数据来源合规证明（包括数据多样性分析报告和知情同意文件）及算法目标函数伦理审查表。在模型训练过程中部署实时监测系统，跟踪记录数据使用轨迹，当检测到敏感数据处理时自动触发伦理复核流程。在临床部署前需通过多中心伦理审查，提供包含 300 例以上真实病例的偏见测试报告，详细说明模型在不同人群组（年龄、性别、地域）的性能差异及修正措施。在模型运行阶段，建立“双轨制”监管：一方面接入医院信息系统实时监控模型决策日志，通过预设的伦理红线指标（如种族敏感性、隐私泄漏风险等）进行自动化扫描；另一方面每月召开伦理听证会，分析典型病例的 AI 决策过程，重点审查争议性案例（如终末期治疗建议）。同时建立全国医疗 AI 伦理案例库，收录各类伦理事件及处置方案，为审查提供参考依据。所有审查过程及结果通过区块链存证，确保全程可追溯且不可篡改。

3. 我国实践路径建议

构建医疗大模型算法备案与溯源系统需依托国

家药品监督管理局与国家卫生健康委员会的联合监管框架，建立三级联动的算法管理体系。国家级平台统一制定备案标准，要求医疗领域大模型开发者提交完整技术文档，包括模型架构、训练数据分布（如年龄、性别、地域的占比）、性能指标及偏见检测报告，并通过区块链技术固化提交内容且确保不可篡改。省级节点负责属地化审核，组织临床专家、数据安全团队及伦理委员会开展多维度评估，重点验证模型在区域代表性数据集上的泛化能力与公平性，审核通过后生成唯一算法备案编码。医疗机构接入层可以部署轻量化溯源终端，实时记录模型调用日志，包括输入数据特征、决策输出及操作者信息，通过国产密码算法加密后同步至省级监管链。在技术实现上，采用“联邦学习+安全多方计算”架构，支持跨机构模型更新时的参数追溯；开发动态审计模块，自动检测算法运行偏差（如群体性能差异超过阈值时触发预警），并关联备案编码生成溯源报告。配套建立算法迭代更新备案机制，要求开发者定期提交再训练数据说明及性能复核结果，确保模型全生命周期可监管。在系统运营中实行“白名单”制度，对通过备案的模型赋予医疗机构采购准入资格，并接入全国医疗 AI 监管信息平台，实现数据互通。

（三）建设临床应用规范

1. 分级分类准入制度

在构建分级分类准入制度时，应依据应用场景的风险等级实施精准管理。对于高风险场景（如辅助诊断、治疗方案生成），严格纳入医疗器械三类监管，要求开发者提交完整的临床验证方案，包括多中心随机对照试验设计、样本量计算（需覆盖不同性别、年龄、种族及疾病分期的代表性人群）及伦理审查批件，试验结果需证明模型灵敏度、特异度及临床一致性达到预设标准（如受试者工作特征曲线下面积（AUC） ≥ 0.90 ），同时提供算法训练数据集的多样性分析报告（包括数据来源、分布均衡性及潜在偏差修正措施）。针对中风险场景（如影像预标注、病历结构化），可采用真实世界证据结合前瞻性观察研究进行验证，要求连续监测至少 6 个月的实际应用数据，并定期提交性能评估报告（如标注准确率 $\geq 95\%$ ）。针对低风险场景（如智能导诊、健康咨询），可以实行备案制管理，并通过

基础安全性测试和功能验证。针对大模型特有风险（如幻觉输出、长尾数据泛化不足），新增强制性测试项目，包括对抗性压力测试（模拟罕见病例输入）和决策可追溯性验证（输出需关联至少3层可解释证据链）。技术审查可由国家药品监督管理局联合多学科专家委员会执行，在技术审查通过后颁发场景限定型注册证书，并纳入国家级医疗AI监管平台进行全生命周期追踪。

2. 人机协同诊疗标准

制定人机协同诊疗标准需明确医生与大模型的权责边界，建立分层次、可追溯的协作机制。在临床决策中，医生始终保留最终裁决权，大模型仅作为辅助工具提供参考建议。在具体实施上，按模型输出置信度划分决策权限：当置信度 $\geq 95\%$ 时（如典型病变识别、常规治疗方案推荐），医生需对关键指标（如病灶定位误差 $\leq 1\text{ mm}$ 、药物禁忌符合率为 100% ）进行复核确认；在置信度为 $80\% \sim 95\%$ 时（如复杂病例鉴别诊断），模型提供差异化建议（如TOP3诊断可能性及支持证据），医生需结合临床经验独立研判；在置信度 $< 80\%$ 时（如罕见病或疑难病例），强制启动人工接管流程，模型仅提供文献检索与数据支持。所有决策过程需通过区块链技术全程留痕，记录模型输入数据、输出建议、医生修正意见及最终决策依据，形成不可篡改的诊疗证据链。同时，建立动态责任分配机制：若医生采纳模型建议导致误诊，开发者承担算法缺陷责任（ 40% ），医生承担临床判断责任（ 60% ）；若医生否决正确建议引发不良后果，责任权重反向分配。医疗机构需配置专职AI督导员，定期评估模型性能并组织医生培训，重点考核对模型局限性（如幻觉输出识别率 $\geq 85\%$ ）及伦理风险的认知。此外，设立人机协同应急小组，确保在模型异常时能无缝切换至传统诊疗模式。

3. 我国实践路径建议

在我国医疗AI监管实践中，应构建“预认证+分类备案”的混合监管模式。参考美国FDA 510(k)认证机制的核心框架，并结合本土医疗体系特点进行监管。针对中低风险医疗大模型应用（如病历质控、影像预标注），建立基于实质等效原则的快速备案通道，允许通过对比已获批同类产品的性能数据（如敏感度、特异度差异 $\leq 5\%$ ）完成注册，同时要求提交真实世界性能监测计划，在省内至少3家

医疗机构进行为期6个月的临床效果追踪。对于高风险场景（如辅助诊断、治疗方案生成），在现有三类医疗器械审批流程中增设算法透明度专项审查，强制公开训练数据分布（包括地域、年龄、性别占比）、偏见测试报告（如不同人群组受试者AUC差异 ≤ 0.05 ）及可解释性验证结果（关键决策因素可追溯率 $\geq 90\%$ ）。建立国家级医疗AI测试数据库，由国家药品监督管理局联合国家卫生健康委员会统一管理，包含覆盖东西部地区的典型病例数据，用于第三方验证模型泛化能力。在省级药监部门设立AI审评中心，配备临床医学、算法工程和伦理审查复合型人才，实施“技术文档+现场检查+临床抽检”三位一体评估。通过认证的模型赋予唯一备案编码，并接入国家医疗AI监管平台实现全流程溯源，同时对基层医疗专用轻量化模型设立优先审评通道，缩短审批周期。

（四）完善法律监管框架

1. 医疗责任认定规则

在医疗责任认定规则制定中，应建立“技术链+临床链”双轨归责体系，依据各方主体对风险的控制能力实施精准问责。在具体实施上，开发者承担算法设计缺陷责任（如训练数据偏差超过阈值、模型可解释性不达标等），需在备案时提交算法安全白皮书并投保产品责任险；医疗机构作为使用方，对模型临床应用的合理性负责（如适应症超范围使用、未执行强制复核流程等），需建立AI诊疗质量控制系统并留存完整的决策日志；医务人员保留最终临床判断权，若未履行专业审慎义务（如盲目采纳低置信度建议、忽视模型风险提示等）则承担相应责任。建立过错推定制度，当发生医疗损害时，可优先推定AI系统相关方存在一定责任，由开发者、医疗机构及医生分别举证自身已履行合规义务，以确保责任划分更加公平合理。引入动态责任比例划分机制，通过区块链存证的诊疗全流程数据（包括模型输入输出、医生操作记录及患者知情同意书），由第三方鉴定机构量化各方责任权重（如算法缺陷占 40% 、临床误判占 60% ）。组建国家级医疗AI纠纷仲裁委员会，整合临床专家、算法工程师和法律人士，开发专用责任溯源工具，自动分析模型决策路径与人工干预节点的关联性。最终形成“技术备案-过程留痕-过错推定-比例担责”

的闭环监管体系。

2. 医疗伦理审查立法

建议制定“医疗人工智能伦理审查办法”，构建“国家-区域-机构”三级联动的伦理审查体系。国家级伦理委员会负责制定统一审查标准，要求高风险医疗AI应用（如辅助诊断、治疗方案推荐）必须提交算法偏见检测报告（包括性别、年龄、地域、种族等维度的性能差异分析，差异率不得超过5%），并公开训练数据来源及分布统计（如基层医疗数据占比不低于20%）。区域级伦理审查中心负责实施动态监管，每季度对已部署模型进行抽样审计，重点检查临床决策一致性（医生与模型诊断符合率 $\geq 90\%$ ）和数据使用合规性（如患者知情同意书覆盖率达到100%）。医疗机构设立AI伦理审查岗，对模型日常使用进行全流程记录，包括每次调用的输入数据、输出结果及医生采纳情况，通过区块链技术实现不可篡改存证。要求企业设立独立的AI伦理委员会，建立透明度分级披露制度，对于高风险应用需向监管部门开放算法核心参数和决策逻辑，向医务人员提供专业版技术文档（含局限性说明），向患者提供通俗版使用说明（含权利告知）。将伦理审查结果与产品注册挂钩，未通过审查的模型不得进入临床应用，已上市产品若发现重大伦理风险（如群体歧视率超过阈值），应立即启动召回程序。

3. 我国实践路径建议

我国医疗AI法律监管应构建“伦理原则+技术标准+分级监管”的本土化框架，明确患者安全至上、技术可控可用、责任可溯可究三大核心原则。在具体实施上，建立医疗AI风险四级分类制度，将辅助诊断、治疗方案推荐等直接涉及患者生命健康的应用列为最高风险等级（A类），实行“临床试验+多中心验证+动态监测”全流程监管；制定医疗数据采集使用规范，要求数据采集遵循“最小够用”原则（如诊断模型仅收集必要临床指标），采用国产密码算法加密存储，并通过联邦学习实现“数据不出院”；实施算法透明度分级披露，A类应用需向国家药品监督管理局报备核心算法逻辑，向医疗机构公开性能指标和局限说明，向患者提供通俗版使用须知；在省级药监部门设立AI审查中心，组建由临床医生、数据科学家和伦理专家构成的复合型评审团队，采用“技术文档审查+临床场景测

试+伦理影响评估”三维评审机制；建立医疗AI安全监测平台，对接医院信息系统实时采集模型运行数据，当检测到性能衰减或伦理风险时自动预警，通过认证的产品可纳入医保采购目录。同时，鼓励三甲医院与AI企业共建联合实验室，在真实医疗环境中验证模型安全性和有效性，形成具有中国特色的“技术研发-临床验证-监管审批”协同创新机制。

四、医疗领域大模型的发展挑战及应对

医疗领域大模型的快速发展在带来技术革新的同时，也暴露出一系列亟待解决的伦理风险与治理难题。从伦理风险识别到治理路径探索，前文已系统分析了数据隐私、算法偏差、主体关系与社会公平等核心风险，并提出了基于“数据-算法-应用-法律”四位一体的治理框架。然而，这些治理路径的有效实施仍需克服技术与政策层面的多重障碍。技术挑战作为治理落地的首要障碍，直接关系到模型的可信度与临床应用的安全性及普适性。例如，数据隐私保护与算法可解释性^[34]不仅是技术问题，更是伦理治理的基础；政策滞后则集中体现为立法和监管更新速度远落后于技术迭代，其成因既包括传统审批流程的刚性，也涉及跨境数据流动中各国标准不一所带来的合规复杂性，进一步加剧了技术落地的难度。未来，通过技术创新与制度完善的协同推进，深入探讨医疗领域大模型在实践中的技术瓶颈及其对治理路径的影响，为后续政策制定与技术优化提供方向，推动医疗领域大模型持续规范健康发展。

（一）技术挑战

1. 大模型的可解释性与信任度提升

医疗领域大模型的可解释性与信任度提升面临显著挑战，其核心在于复杂神经网络架构的“黑箱”特性导致决策过程难以透明化。当前，大模型基于海量参数和多层非线性变换生成预测结果，但内部推理逻辑缺乏直观的医学意义关联。例如，在医学影像诊断中，模型可能依赖与临床经验不符的隐式特征（如图像背景噪声或无关纹理）进行判断，而注意力机制生成的热图虽能标注疑似病灶区域，却无法解释其与病理学标准的对应关系（如恶

性结节的毛刺征或分叶征的医学依据)。为解决这一问题,研究者正探索多模态可解释技术,如集成梯度和概念激活向量,试图将模型决策映射到医学概念(如特定生物标志物或解剖特征),但现有方法在跨模态数据(如影像与基因组数据联合分析)中的解释一致性不足,且无法覆盖模型的全链路推理过程。此外,可解释性需求存在场景差异性,如急诊决策需实时可视化关键证据链,而科研场景则要求细粒度参数分析,这对通用解释框架的设计提出了更高要求。

2. 多模态数据融合的隐私保护难题

医疗领域多模态数据融合的隐私保护面临严峻挑战,主要体现在跨模态关联导致的敏感信息泄漏风险。医疗大模型通常需要整合EHR、医学影像、基因组数据和可穿戴设备信息等多源异构数据,但在融合过程中,不同模态间的隐含关联可能暴露患者隐私。例如,基因组数据与面部特征的关联分析可能推断出患者的遗传病风险,而医学影像的元数据(如DICOM头文件)结合诊疗记录可能重识别患者身份。现有隐私保护技术如联邦学习虽能实现分布式训练,但在跨模态特征对齐时,梯度更新仍可能泄露数据分布特征;差分隐私虽能添加噪声保护个体数据,却会显著降低多模态关联分析的精度。更复杂的是,不同数据类型对隐私保护的要求各异:基因组数据需终身保护,而临床指标可能只需短期匿名化。当前研究正探索分层隐私保护架构:在数据层采用同态加密处理敏感字段(如基因位点),在特征层使用MPC进行跨模态匹配,在模型层则通过联邦学习与知识蒸馏相结合的方式提取共享知识。

(二) 政策挑战

1. 跨国数据流动与伦理标准化困境

医疗领域大模型的跨国数据流动与伦理标准化建设面临双重困境,主要表现为国际法律框架冲突与伦理审查体系碎片化。在数据跨境流动方面,各国监管要求存在根本性差异:欧盟GDPR实行严格的“充分性保护”原则,要求数据接收方达到欧盟隐私标准;《中华人民共和国个人信息保护法》强调数据本地化存储,重要医疗数据出境需通过安全评估;而美国《健康保险可携性与责任法案》(HIPAA)允许数据流向签署“隐私盾”协议的国

家,但缺乏统一的国际互认机制。在伦理审查层面,各国对医疗AI的伦理标准同样分化:欧盟《人工智能法案》要求高风险医疗AI提供全链条算法透明度,美国FDA更关注临床有效性证据,我国则强调数据主权与国家安全,同一医疗领域大模型在不同国家可能面临完全不同的伦理约束。当前,国际组织如世界卫生组织发布了《世界卫生组织卫生健康领域人工智能伦理与治理指南》^[35],但缺乏强制力;部分企业尝试通过“数据飞地”(如设立跨境安全计算环境)规避法律冲突,却面临技术可行性与成本收益的质疑。

2. 全周期监管滞后与技术迭代脱节

医疗AI领域的监管滞后问题日益凸显,主要表现为传统审批框架难以匹配大模型快速迭代的技术特性。当前医疗AI监管普遍采用“先审批后应用”的静态模式,平均审批周期长达12~18个月,而大模型技术如GPT系列已实现每年一代的迭代速度,导致获批产品在上市时技术已然落后。美国FDA虽已试点“预认证计划”来加快AI医疗软件的审批,但仍局限于小规模模型更新;欧盟《人工智能法案》提出了适应性监管概念,但具体实施路径尚不明确。我国正在探索的“算法备案+动态监测”机制虽具有一定前瞻性,但对模型迭代的实时跟踪能力仍显不足。解决以上困境需要构建“监管沙盒+实时审计”的新型治理范式:通过TEE实现监管机构对模型参数的实时验证,利用区块链技术记录每一次迭代变更,并建立基于风险的动态分级监管机制。同时,需要开发自动化合规测试工具,将伦理审查指标(如公平性、可解释性)嵌入模型开发流水线,实现从“事后监管”向“嵌入式治理”的转变。

(三) 应对建议

1. 探索基于区块链医疗数据确权

区块链技术在医疗数据确权领域的应用为解决当前医疗数据共享中的信任缺失问题提供了创新路径。基于区块链的医疗数据管理系统能够实现患者数据的去中心化确权与精细化管理,通过智能合约自动执行数据访问控制策略,确保患者在保留数据所有权的前提下实现可控共享。具体而言,该系统可以将患者的电子病历、影像数据、基因组信息等关键医疗数据以哈希值形式存储于区块链,原始数

据则加密保存在医疗机构本地，既满足《中华人民共和国个人信息保护法》的数据本地化要求，又能通过分布式账本实现跨机构数据使用的全程追溯。值得注意的是，当前医疗区块链系统仍面临吞吐量低、存储成本高、与现有医疗信息系统集成困难等挑战。未来发展方向包括：构建医疗数据要素市场，通过通证经济激励数据贡献；开发轻量级联盟链解决方案，降低基层医疗机构接入门槛；建立联邦学习融合架构，实现“数据不动模型动”的安全协作模式等，从根本上重塑医疗数据治理范式，为医疗领域大模型发展提供合规、高效的数据支撑。

2. 开发轻量化模型普惠基层医疗

轻量化医疗领域大模型的开发是实现 AI 技术普惠基层医疗的关键突破口。当前研究主要从 3 个维度推进模型轻量化：在模型架构方面，通过知识蒸馏技术将大型模型的核心能力迁移至小型网络（如将百亿参数模型压缩至千万级），并采用自适应计算机制实现推理时动态调整计算量；在数据利用层面，开发基于小样本学习的训练范式，利用迁移学习缓解基层医疗数据不足的问题，同时通过数据增强技术提升模型在基层常见病上的识别精度；在部署优化上，创新边缘计算架构，支持模型在低配图形处理器甚至移动设备上实时运行，并开发断网续推功能以应对网络波动。值得注意的是，轻量化不应简单等同于性能降级。未来发展方向包括：开发“云-边-端”协同推理框架，实现资源动态调配；构建基层医疗专用模型库，覆盖常见病、多发病的筛查需求；建立模型自适应进化机制，使系统能够根据基层数据反馈持续优化等，有效缩小城乡医疗 AI 应用的“数字鸿沟”，为分级诊疗制度的落实提供智能化支撑。

3. 构建“政产学研医”协同生态系统

构建“政产学研医”协同生态系统是推动医疗领域大模型规范健康发展的关键路径。政府部门应主导制定医疗领域大模型伦理治理白皮书，明确数据主权归属和跨境流动规则，同时设立国家级医疗 AI 伦理委员会，建立覆盖模型研发、测试、应用全周期的动态“监管沙盒”。医疗机构需与科技企业共建联合创新中心，在真实医疗场景中验证模型的有效性和安全性，并建立临床反馈快速响应机制。高校和科研院所则应打破学科壁垒，开设“医学+AI+伦理”交叉学科，培养既懂临床需求又掌握 AI 技

术的复合型人才，同时加强在联邦学习、可解释 AI 等关键技术领域的原始创新。产业界需要成立医疗 AI 联盟，制定行业自律公约，共享安全治理经验，并通过建立医疗 AI 伦理风险共担基金来化解创新风险。特别重要的是构建多方参与的评价体系，包括由临床专家主导的效用评估、由患者代表参与的体验反馈、由技术专家执行的安全审计，最终形成技术创新与伦理治理良性互动的可持续发展格局，为“健康中国”战略提供智能化支撑^[3]。

五、结语

医疗领域大模型的伦理治理是一项需要技术创新与制度保障协同推进的系统工程^[6]。作为医疗数字化转型的核心驱动力，大模型在提升诊疗效率的同时，其特有的数据依赖性、算法复杂性和应用广泛性也带来了前所未有的伦理挑战^[6]。当前治理路径需重点突破 4 个维度的协同：在数据层面，建立基于隐私计算的“数据可用不可见”共享机制，实现敏感医疗数据的价值释放与隐私保护平衡；在算法层面，开发医疗专用的可解释性技术，通过临床知识引导的注意力机制和决策路径可视化，破解“黑箱”困境；在应用层面，构建分级分类的准入体系，针对辅助诊断、健康管理等不同风险等级场景实施差异化监管；在法律层面，完善动态责任认定框架，采用区块链存证技术实现诊疗全流程可追溯。值得注意的是，医疗领域大模型的伦理治理不能简单照搬通用 AI 的治理模式，必须充分考虑医疗行业的特殊性，如患者隐私的敏感性、诊疗决策的不可逆性以及医疗资源的公平性诉求。未来需要建立“技术-制度-文化”三位一体的治理生态：技术上持续优化联邦学习、同态加密等隐私保护方案；制度上完善算法备案、伦理审查等监管工具；文化上培育“负责任创新”的行业共识，真正实现医疗领域大模型“向善而行”的发展愿景，使其成为推动优质医疗资源普惠共享、助力“健康中国”战略落地的关键支撑。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: July 19, 2025; Revised date: August 26, 2025

Corresponding author: Wang Junpu is an associate professor from Xiangya Hospital / Xiangya School of Basic Medicine, Central South

University. His major research fields include pathology and digital intelligence pathology. E-mail: wang-jp2013@csu.edu.cn

Funding project: Furong Laboratory Project (2024PT5111)

参考文献

- [1] Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine [J]. *Nature Medicine*, 2022, 28(1): 31–38.
- [2] 中国信息通信研究院知识产权与创新发展中心, 中国信息通信研究院科技伦理研究中心. 人工智能伦理治理研究报告 [R]. 北京: 中国信息通信研究院, 2023. Intellectual Property and Innovation Development Center of China Academy of Information and Communications Technology, Science and Technology Ethics Research Center of China Academy of Information and Communications Technology. Research report on AI ethics governance [R]. Beijing: China Academy of Information and Communications Technology, 2023.
- [3] 余玉刚, 王耀刚, 江志斌, 等. 智慧健康医疗管理研究热点分析 [J]. *管理科学学报*, 2021, 24(8): 58–66. Yu Y G, Wang Y G, Jiang Z B, et al. Analysis of research hotspots of intelligent health care management [J]. *Journal of Management Sciences in China*, 2021, 24(8): 58–66.
- [4] Maddox T M, Embí P, Gerhart J, et al. Generative AI in medicine—Evaluating progress and challenges [J]. *New England Journal of Medicine*, 2025, 392(24): 2479–2483.
- [5] 郭华源, 刘盼, 卢若谷, 等. 人工智能大模型医学应用研究 [J]. *中国科学(生命科学)*, 2024, 54(3): 482–506. Guo H Y, Liu P, Lu R G, et al. Research on medical application of artificial intelligence large model [J]. *Science in China (Vita)*, 2024, 54(3): 482–506.
- [6] Webster P. Six ways large language models are changing health-care [J]. *Nature Medicine*, 2023, 29(12): 2969–2971.
- [7] Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medical artificial intelligence [J]. *Nature*, 2023, 616(7956): 259–265.
- [8] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge [J]. *Nature*, 2023, 620(7972): 172–180.
- [9] Xiang J X, Wang X Y, Zhang X M, et al. A vision-language foundation model for precision oncology [J]. *Nature*, 2025, 638(8051): 769–778.
- [10] 陈晓红, 刘浏, 牛雅娟, 等. 数智病理平台构建及服务模式研究 [J]. *中国工程科学*, 2025, 27(2): 304–314. Chen X H, Liu L, Niu Y J, et al. Digital intelligence pathology platform and its service pattern [J]. *Strategic Study of CAE*, 2025, 27(2): 304–314.
- [11] 国家卫生健康委员会办公厅, 国家中医药管理局综合司, 国家疾病预防控制中心综合司. 卫生健康行业人工智能应用场景参考指引 [EB/OL]. (2024-11-06)[2025-07-17]. <https://www.nhc.gov.cn/wjw/c100175/202411/5bcb3c4edd064e31ac5d279caf5830f4.shtml>. General Office of the National Health Commission, Comprehensive Department of the National Administration of Traditional Chinese Medicine, Comprehensive Department of the National Disease Control and Prevention Administration. Reference guidelines for artificial intelligence application scenarios in the health industry [EB/OL]. (2024-11-06)[2025-07-17]. <https://www.nhc.gov.cn/wjw/c100175/202411/5bcb3c4edd064e31ac5d279caf5830f4.shtml>.
- [12] 陈晓红, 刘浏, 袁依格, 等. 医疗大模型技术及应用发展研究 [J]. *中国工程科学*, 2024, 26(6): 77–88. Chen X H, Liu L, Yuan Y G, et al. Technology and application development of medical foundation model [J]. *Strategic Study of CAE*, 2024, 26(6): 77–88.
- [13] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine [J]. *Nature Medicine*, 2023, 29(8): 1930–1940.
- [14] Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations [J]. *Science*, 2019, 366(6464): 447–453.
- [15] Rocher L, Hendrickx J M, de Montjoye Y A. Estimating the success of re-identifications in incomplete datasets using generative models [J]. *Nature Communications*, 2019, 10: 3069.
- [16] 施锦诚, 王国豫, 王迎春. ESG视角下人工智能大模型风险识别与治理模型 [J]. *中国科学院院刊*, 2024, 39(11): 1845–1859. Shi J C, Wang G Y, Wang Y C. Artificial intelligence foundation model risk identification and governance model from the ESG perspective [J]. *Bulletin of Chinese Academy of Sciences*, 2024, 39(11): 1845–1859.
- [17] 中华人民共和国国务院. 新一代人工智能发展规划 [EB/OL]. (2017-07-08)[2025-07-17]. https://www.gov.cn/gongbao/content/2017/content_5216427.htm. State Council of the People's Republic of China. Next generation artificial intelligence development plan [EB/OL]. (2017-07-08)[2025-07-17]. https://www.gov.cn/gongbao/content/2017/content_5216427.htm.
- [18] 中华人民共和国国家互联网信息办公室, 中华人民共和国国家发展和改革委员会, 中华人民共和国教育部, 等. 生成式人工智能服务管理暂行办法 [EB/OL]. (2023-07-10)[2025-07-18]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm. Cyberspace Administration of China, National Development and Reform Commission, Ministry of Education of the People's Republic of China, et al. Interim measures for the management of generative artificial intelligence services [EB/OL]. (2023-07-10)[2025-07-18]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.
- [19] 李强. 政府工作报告 [EB/OL]. (2025-03-12)[2025-07-18]. https://www.gov.cn/yaowen/liebiao/202503/content_7013163.htm. Li Q. Government work report [EB/OL]. (2025-03-12)[2025-07-18]. https://www.gov.cn/yaowen/liebiao/202503/content_7013163.htm.
- [20] 国家药品监督管理局. 人工智能医用软件产品分类界定指导原则 [EB/OL]. (2021-07-01)[2025-07-18]. <https://www.nmpa.gov.cn/xxgk/ggtg/ylqxggtg/ylqxqtggtg/20210708111147171.html>. National Medical Products Administration. Guidelines for the classification and definition of artificial intelligence medical software products [EB/OL]. (2021-07-01)[2025-07-18]. <https://www.nmpa.gov.cn/xxgk/ggtg/ylqxggtg/ylqxqtggtg/20210708111147171.html>.
- [21] Zhang Z Q, Yan C, Malin B A. Membership inference attacks against synthetic health data [J]. *Journal of Biomedical Informatics*, 2022, 125: 103977.

- [22] 杨善林, 丁帅, 顾东晓, 等. 医疗健康大数据驱动的知识发现与知识服务方法 [J]. 管理世界, 2022, 38(1): 219–229.
Yang S L, Ding S, Gu D X, et al. Healthcare big data driven knowledge discovery and knowledge service approach [J]. Journal of Management World, 2022, 38(1): 219–229.
- [23] 罗映宇, 朱国玮, 钱无忌, 等. 人工智能时代的算法厌恶: 研究框架与未来展望 [J]. 管理世界, 2023, 39(10): 205–233.
Luo Y Y, Zhu G W, Qian W J, et al. Algorithm aversion in the era of artificial intelligence: Research framework and future agenda [J]. Journal of Management World, 2023, 39(10): 205–233.
- [24] McCradden M D, Joshi S, Mazwi M, et al. Ethical limitations of algorithmic fairness solutions in health care machine learning [J]. The Lancet Digital Health, 2020, 2(5): 221–223.
- [25] Benjamin R. Assessing risk, automating racism [J]. Science, 2019, 366(6464): 421–422.
- [26] 赵玲玲, 郭遥. 智能医疗机器人伦理风险: 类型、成因与防控策略 [J]. 医学与哲学, 2023, 44(12): 35–39.
Zhao L L, Guo Y. Ethical risks of intelligent medical robots: Types, causes, prevention and control strategies [J]. Medicine & Philosophy, 2023, 44(12): 35–39.
- [27] Maliha G, Gerke S, Cohen I G, et al. Artificial intelligence and liability in medicine: Balancing safety and innovation [J]. The Milbank Quarterly, 2021, 99(3): 629–647.
- [28] Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments [J]. Scientific Reports, 2023, 13: 16492.
- [29] Cordeiro J V. Digital technologies and data science as health enablers: An outline of appealing promises and compelling ethical, legal, and social challenges [J]. Frontiers in Medicine, 2021, 8: 647897.
- [30] 唐露源, 谢士尧, 徐源. 大模型产业创新生态系统构建及风险识别研究 [J]. 科学学研究, 2025, 43(9): 1972–1981.
Tang L Y, Xie S Y, Xu Y. Research on the construction and risk identification of large language model industrial innovation ecosystem [J]. Studies in Science of Science, 2025, 43(9): 1972–1981.
- [31] Cao K, Xia Y D, Yao J W, et al. Large-scale pancreatic cancer detection *via* non-contrast CT and deep learning [J]. Nature Medicine, 2023, 29(12): 3033–3043.
- [32] 汪琛. 医疗人工智能伦理治理的问题、困境与求解 [J]. 科学学研究, 2025, 43(2): 414–422.
Wang C. Research on governance over ethics of AI4health/medicine: Problems, dilemmas and solutions [J]. Studies in Science of Science, 2025, 43(2): 414–422.
- [33] 李润生. 论医疗人工智能的法律规制——从近期方案到远期设想 [J]. 行政法学研究, 2020 (4): 46–57.
Li R S. On the legal regulation of medical artificial intelligence—From short-term plan to long-term plan [J]. Administrative Law Review, 2020 (4): 46–57.
- [34] Imrie F, Davis R, van der Schaar M. Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare [J]. Nature Machine Intelligence, 2023, 5(8): 824–829.
- [35] World Health Organization. Ethics and governance of artificial intelligence for health [EB/OL]. (2021-06-28)[2025-07-18]. <https://www.who.int/publications/i/item/9789240029200>.
- [36] 陈晓红, 刘晓亮, 王俊普, 等. 从 AI 驱动到元宇宙赋能: 病理诊断管理模式创新与扩展 [J/OL]. 中国管理科学, 1–14[2025-09-17]. <https://doi.org/10.16381/j.cnki.issn1003-207x.2023.1211>.
Chen X H, Liu X L, Wang J P, et al. From AI-driven to metaverse-empowered: Innovation and expansion of pathological diagnosis management models [J/OL]. Chinese Journal of Management Science, 1–14[2025-09-17]. <https://doi.org/10.16381/j.cnki.issn1003-207x.2023.1211>.