

DOI: 10.61189/922448aqtvt

· 方法学研究 ·

医学人工智能体的风险适配型全生命周期框架： MAAM - MAAE

田卫彬¹, 周建¹, 余津津¹, 杨达伟^{1,2,3,4}, 白春学^{1,2,3,4*}

1. 复旦大学附属中山医院呼吸与危重症医学科, 上海 200032

2. 复旦大学附属中山医院AI+肺癌防治中心, 上海 200032

3. 上海呼吸物联网医学工程技术研究中心, 上海 200032

4. 上海市呼吸病研究所, 上海 200032

[摘要] 医学人工智能正从单输出模型演进为能够保持上下文连贯、执行多步骤规划、调用工具并发起或草拟临床行动的智能体。与传统模型相比,智能体的风险不再仅由单次输出的正确与否来度量,更取决于它如何进行任务规划、权限调用、记忆写入、工具选用,以及其与临床团队之间的协同。错误一旦沿上述任一环节传递,便有可能从建议层面的偏差转化为行动层面的后果。因此,医学智能体进入临床前的评估,不能仅依赖模型自身的性能指标,而必须覆盖人-智能体- workflow 构成的完整系统。本文关注的核心问题是:依据医学智能体行动后果的严重程度,究竟需要何种层级的系统证据,才能判定其是否具备安全进入并持续运行于临床的条件。本研究通过叙述性综述方式整合了医学人工智能、人因工程、安全工程、网络安全、实施科学、卫生经济学、报告规范及监管科学等跨学科文献,构建了MAAM-MAAE这一风险适配型概念方法学。我们将现有标准映射至智能体特有的属性与风险,并将尚未解决的要求梳理为相互衔接的研发与评价循环。需说明的是,本研究采用目的性而非系统性证据整合,且尚未开展正式的专家共识程序或前瞻性验证。医学智能体架构与成熟度(medical agent architecture and maturation, MAAM)将研发活动划分为4个领域:临床使命、系统架构、人机协作(含行动权限),以及安全与成熟度保障。在评估层面,MAAM引入医学智能体评估与评价(medical agent assessment and evaluation, MAAE)从能力、安全、人-智能体与临床表现、以及卫生系统与全生命周期价值4个层级面组织证据。评价单元为完整的人-智能体-workflow系统。4个风险层级与五个阶段关口——关口设置依据潜在伤害、行动与患者的物理接近程度、自主程度、错误可检测性及可逆性,以此使监督强度与所需证据水平相匹配。此外本文提出以临床任务成功率(clinical task success, CTS)和临床后果性不安全行动率(clinically consequential unsafe action rate, CCUAR)作为双重锚定指标,但二者的测量属性及决策阈值尚需实证研究加以检验。MAAM-MAAE将临床授权从一次性模型验证决策,转变为面向特定版本的全生命周期契约。该框架通过纳入行动、纵向运行、变更控制和退出机制,对现有报告及监管框架形成补充。MAAM-MAAE是一项可检验的提案,而非已经验证的核查表、报告指南或临床有效性声明。

[关键词] 医学智能体;智能体人工智能;临床评价;人-医智能体协作;患者安全;全生命周期治理;方法学

[中图分类号] R-3 **[文献标志码]** A

A risk-adaptive full lifecycle framework for medical AI agents: MAAM - MAAE

Tian Weibin¹, Zhou Jian¹, Yu Jinjin¹, Yang Dawei^{1,2,3,4}, Bai Chunxue^{1,2,3,4*}

1. Department of Respiratory and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China

2. AI+Lung Cancer Prevention and Treatment Center, Zhongshan Hospital, Fudan University, Shanghai 200032, China

3. Shanghai Respiratory Internet of Things Medical Engineering Technology Research Center, Shanghai 200032, China

4. Shanghai Institute of Respiratory Diseases, Shanghai 200032, China

[Abstract] Medical artificial intelligence is evolving from single-output models into agents capable of maintaining contextual

[收稿日期] 2026-07-10

[接受日期] 2026-09-05

[基金项目] 四大慢病重大专项(2024ZD0529300),国家自然科学基金(82570143、82170110),上海市科学技术委员会项目基金(20DZ2254400、21DZ2200600、20DZ2261200),上海市市级科技重大专项(ZD2021CY001)和上海市临床重点专科建设项目(shslczdk02201). Supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0529300), the National Natural Science Foundation of China (82570143 and 82170110), the Science and Technology Commission of Shanghai Municipality (20DZ2254400, 21DZ2200600, and 20DZ2261200), the Shanghai Municipal Science and Technology Major Project (ZD2021CY001), and the Shanghai Municipal Key Clinical Specialty (shslczdk02201).

[作者简介] 田卫彬,硕士生;E-mail: 26111020031@m.fudan.edu.cn

*通信作者(Corresponding author). 白春学,博士,主任医师、教授;E-mail: cxbai@fudan.edu.cn

coherence, performing multi-step planning, invoking tools, and initiating or drafting clinical actions. Compared with traditional models, the risk of agents is no longer measured solely by whether a single output is correct; it depends more on how they perform task planning, invoke permissions, write to memory, select tools, and coordinate with the clinical team. Once an error propagates through any of these links, it may translate from a deviation at the level of recommendation into consequences at the level of action. Therefore, evaluation before medical agents enter clinical settings cannot rely only on the model's own performance metrics; it must cover the complete system formed by the human-agent-workflow. The core question of this paper is: according to the severity of the consequences of medical agent actions, what level of system evidence is required to determine whether an agent can safely enter and continuously operate in clinical settings. Through a narrative review, this study integrates interdisciplinary literature from medical artificial intelligence, human factors engineering, safety engineering, cybersecurity, implementation science, health economics, reporting standards, and regulatory science, and constructs MAAM-MAAE, a risk-adaptive conceptual methodology. We map existing standards to agent-specific attributes and risks and organize unresolved requirements into interconnected R&D and evaluation cycles. It should be noted that this study uses purposive rather than systematic evidence synthesis and has not conducted a formal expert consensus process or prospective validation. Medical Agent Architecture and Maturation (MAAM) divides R&D activities into four domains: clinical mission, system architecture, human-agent collaboration (including action authority), and safety and maturity assurance. At the assessment level, MAAM introduces Medical Agent Assessment and Evaluation (MAAE) to organize evidence across four hierarchical dimensions: capability, safety, human-agent and clinical performance, and health system and full-lifecycle value. The unit of evaluation is the complete human-agent-workflow system. Four risk tiers and five stage gates—the gates are set according to potential harm, the physical proximity of actions to patients, the degree of autonomy, error detectability, and reversibility—so that supervision intensity matches the required level of evidence. In addition, this paper proposes clinical task success (CTS) and clinically consequential unsafe action rate (CCUAR) as dual anchoring metrics; however, their measurement properties and decision thresholds still require empirical testing. MAAM-MAAE transforms clinical authorization from a one-time model validation decision into a version-specific full-lifecycle contract. By incorporating actions, longitudinal operation, change control, and exit mechanisms, this framework complements existing reporting and regulatory frameworks. MAAM-MAAE is a testable proposal, not a validated checklist, reporting guideline, or claim of clinical effectiveness.

[Key Words] medical agents; agentic artificial intelligence; clinical evaluation; human-medical agent collaboration; patient safety; full-lifecycle governance; methodology

医学人工智能正从单一预测或回答的系统,向目标导向型智能体演进,它能够长期观察患者、检索病历和外部知识、保持记忆、分解任务、制定计划、调用工具,并根据反馈迭代行为。因此,具有临床意义的干预已不再等同于单个基础模型,而是由模型、数据、知识、记忆、工具、权限、界面、人类监督及工作流共同构成的社会技术系统^[1-8]。近期综述显示,尽管智能体架构正在医疗领域迅速扩展,但其定义和评价框架仍存在不确定性^[9-10]。

以模型为中心的评价不足以代表该系统。在接近真实临床的模拟环境中,大语言模型在诊断准确性、指南依从性、检查解读、指令执行及稳健性等方面均存有不足^[11]。一项预注册随机研究($n=1298$)^[12]显示,模型独立使用时,可在 94.9% 的情境中识别相关疾病,但参与者借助同一模型后,识别率不足 34.5%;且其处置决策并不优于使用常规信息来源者。相反,在受控条件下,AMIE 在诊断对话、多模态远程会诊和纵向疾病管理任务中展现较好性能;另有研究^[13-17]表明,大语言模型辅助可减少复杂心脏病学病例中的重要临床遗漏,随机试验也

提示 GPT-4 辅助或能改善医生对患者照护任务的完成质量。综合来看,模型的独立能力并不能决定人一智能体系统的整体表现。

现有医学人工智能评价体系中,报告指南已覆盖人工智能试验方案与报告、早期临床评价、诊断准确性、预测模型、大语言模型、医学影像人工智能及可信部署等多个方向^[18-26];卫生经济评价标准涉及成本和价值^[27-28];共识与指南制定框架关注方法学透明度^[29-30];分阶段临床评价文献涉及转化和试验^[31-33];风险管理、监管和网络安全框架则覆盖全生命周期控制^[34-42]。这些文件均不可或缺,但尚无单一框架能将持久记忆、多步骤规划、工具介导行动、权限边界、错误恢复、多智能体协同、版本依赖及纵向运行整合为一条可审计的证据链。因此,我们提出:在医学智能体获准进入临床使用之前,需要哪些与风险相匹配的证据;部署后,又需要哪些持续证据来证明其仍可安全运行?

为回答上述问题,本文提出 MAAM-MAAE 框架。其核心原则是:随着智能体行动的潜在后果愈发严重——即其与患者的接近度越高、自主性越

强、伤害发生前错误可检测性越低、可逆性越差——设计控制、人类监督层级以及临床证据强度均应同步提高。MAAM-MAAE框架的临床目标并非追求自主程度的最大化或基准测试的最优表现,而是证实完整的人—智能体— workflow 系统相较现行实践是否具有净获益。

1 方法

1.1 设计与适用范围 MAAM-MAAE 是基于叙述性证据整合形成的概念性方法学,适用于至少具备以下三项特征之一的目标导向型医学人工智能系统:跨步骤规划、与工具或环境交互,或依据反馈调整自身行为。其适用对象包括单智能体和多智能体系统、面向临床人员的辅助智能体、面向患者的智能体及 workflow 智能体。术语的泛化会掩盖智能体属性所带来的风险,因此对于普通聊天机器人、单一预测模型和固定规则自动化系统,除非其行为符合上述界定范围,否则不应被称为智能体。

该框架可用于指导研发方案、伦理审查、采购评估、早期临床研究、监管评价、医疗机构上线决策及部署后再认证,但不能替代特定司法辖区的法规、临床治理规范或针对特定研究设计的报告指南。每项评价均应前瞻性明确其主要推断单元:基础模型、编排层、完整智能体、智能体辅助的临床人员、患者—智能体交互或完整的人—智能体— workflow 系统。若涉及多智能体系统,则还需明确各组件角色、通信方式、共享记忆、仲裁机制和相关性失效。

1.2 证据整合与框架构建 采用目的性整合策略梳理了医学人工智能、人因工程、安全工程、网络安全、实施科学、卫生经济学、报告指南和医疗器械监管等领域的文献。在此基础上,将主要标准映射为3个核心问题:它所回答的问题是什么、哪些智能体特异性风险尚未充分规定以及 MAAM-MAAE 增加何种要求(表1)。由此形成的解释性矩阵进一步构建了4个研发领域、4个评价层级、风险—权限矩阵、5个转化关口、双重锚定指标(获益与安全)以及最低报告数据集。

鉴于本次文献整合并非系统性过程,我们未估算效应量、未评价证据确定性,也不主张覆盖全部相关文献。框架各组成部分均基于映射所得缺口提出,属于规范性建议,而非实证研究发现。若要获得正式指南地位,还需开展注册的证据综述、透明的共识程序、患者和公众参与、利益冲突管理、公开征求意见及前瞻性检验;其制定与报告宜参照

ACCORD、AGREE II 等相关指南^[29-30]。

2 MAAM-MAAE 框架

智能体的能力改变了错误抵达患者的机制。一个错误输出可能演变为有缺陷的计划,该计划可能选择错误工具、应用错误参数、写入受污染的记忆、触发另一智能体或促使人类临床人员采信误导性建议。真实临床中的危险,往往不是来自于模型单次的错误应答,而是错误沿着工具、权限和 workflow 持续传导。因此,最终伤害不仅取决于模型本身的错误,更取决于架构、权限、验证、界面和 workflow,它们决定了失效轨迹是被放大还是被截断。MAAM 规定了临床暴露前应具备的控制措施;MAAE 则在逐步接近真实临床的条件下检验这些控制是否有效。评价结果重新反馈至研发环节,使该框架形成两个相互嵌套的学习循环,而非单向流水线(图1)。

3 MAAM:医学智能体架构与成熟度

3.1 临床使命 研发应始于尚未满足的临床需求,而不是为现有模型寻找适配的应用场景。预期用途说明应明确目标人群、疾病阶段、应用场所、使用者、任务起止点、在现行路径中的位置、预期获益、可接受的残余风险、排除用途及失效后的替代路径。拟解决的临床问题应具有可行动性,例如在提高指南一致性的同时减少肺结节检出后的失访,而不是泛泛提出“提高效率”。若固定规则、传统预测模型或 workflow 重构已足以更低风险解决问题,则引入智能体只会增加不必要的复杂性。

预期用途还应包括反事实基线:在无智能体介入时由谁执行任务、需要多长时间、会发生哪些错误以及患者和工作人员需承担何种负担。该基线用于确定目标产品特征和最小临床重要差异。在价值主张层面,则需从患者、专业人员和系统3个层面同时评价获益、安全、时间、工作负担、公平性和资源消耗。

3.2 系统架构 最低功能闭环涵盖“感知—推理—规划—行动—验证—适应”六个环节。感知整合病历、对话记录、影像资料、生理信息和环境状态;受治理的知识连接指南、处方集、政策和患者特异性信息;记忆应区分工作记忆、患者纵向记忆和系统经验;规划负责分解并确定任务优先级;工具执行检索、计算、下单、预约、消息发送或设备操作;验证用于检查输出;适应仅可在授权边界内改变后续行为^[43-47]。

表 1 现有卫生人工智能标准的方法学缺口及 MAAM-MAAE 的补充贡献

标准或框架	主要对象与阶段	所回答的问题	智能体特有缺口	MAAM-MAAE 的补充
TRIPOD+AI /PROBAST+AI	预测模型研发、验证及偏倚风险	数据、建模、性能、适用性及开放科学	多步骤任务、工具链、记忆及行动后果	轨迹审计、长时程稳定性、行动层级风险及特定版本评价
STARD-AI	采用人工智能的诊断准确性研究	数据集、待评价试验、参考标准及公平性	诊断后的规划、执行、恢复和升级	诊断—计划—行动闭环及具有临床后果的错误
SPIRIT-AI/CONSORT-AI	人工智能干预试验方案与报告	预期用途、干预描述、输入、输出及错误分析	持续更新、权限变更、接管和越权	智能体版本、AUL、HITL、停止规则及考虑更新的估计目标
DECIDE-AI	早期、小规模临床评价	工作流、人因、可行性及安全性	长时程运行、多智能体级联及部署后漂移	静默部署、压力与恢复测试及持续监测
TRIPOD-LLM	采用生物医学大语言模型的研究	提示词、模型访问、数据泄漏及可重复性	工具执行、持久记忆及目标导向行为	架构档案、工具契约、记忆边界、规划及验证
CLAIM	医学影像人工智能	影像数据、参考标准、训练及评价透明度	跨模态整合及下游行动	缺失或冲突模态测试及行动链追踪
CHEERS/CHEERS-AI	经济学评价	评价视角、成本、结局、不确定性及公平性	推理、验证、版本替换及外部性成本	每项成功任务的成本、接管与更新成本及环境足迹
FUTURE-AI	可信、可部署的卫生人工智能	公平性、普适性、可追溯性、可用性、稳健性及可解释性	自主行动和动态权限的操作化不足	风险、自主程度、监督强度与证据水平四者之间的匹配
WHO/NIST AI RMF	伦理治理与风险管理	人权、治理以及风险识别、测量和管理	后果分级及临床行动的操作关口	双重锚定结局、4 个风险等级及 5 个阶段关口
FDA TPLC/PCCP;IMDRF SaMD	医疗器械产品全生命周期	风险分类、临床评价、变更及持续有效性	通用模型供应链及提示词或工具频繁变更	组件层级版本谱系、变更影响分析、回滚及再验证

AUL: 行动权限等级; HITL: 人在环; PCCP: 预定变更控制计划; SaMD: 医疗器械软件; TPLC: 产品全生命周期。

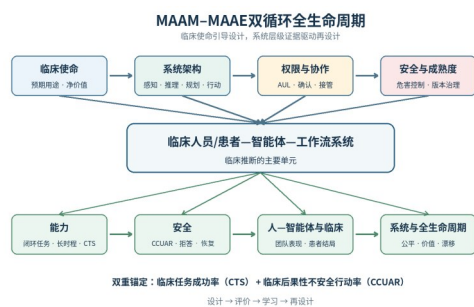


图 1 MAAM - MAAE 双循环全生命周期及其系统层级评价单元

安全控制应模块化。工具调用应通过契约进行约束,明确输入、输出、超时、幂等性和权限。检索应优先采用可追溯、带时间戳的来源。写入记忆前应核验身份、时间、来源和临床状态。高风险任务应将规划与执行分离,并设置独立验证者。当多个智能体共享模型、数据、提示词或失效模式时,多数表决并不能提供保护;有意义的冗余要求模型、来源、角色或验证机制具有异质性。

智能体架构档案作为核心研发产物,内容包括组件清单、数据流图、模型及提示词版本、知识库时间戳、记忆规则、工具清单、权限矩阵、日志字段、外部依赖、网络边界和回滚程序。对于闭源系统,还应记录可获得的模型卡、数据保留条款、漂移检测方法以及供应商停止服务的应急预案。

3.3 行动权限与人类协作 “临床医生仍负最终责任”这一声明本身并不是安全控制。监督机制必须明确:由谁设定目标、复核输入、批准高风险行动、拥有否决权、接收升级信息、接管系统和提供替代服务,以及上述操作应在多长时间内完成。本文提出四级行动权限(action authority level, AUL): A1 为信息处理; A2 为给出建议; A3 为起草行动(未经授权不得生效); A4 为在预先授权范围内实施有界执行。AUL 界定系统的可操作边界,而风险则描述系统出错后的后果,二者必须分别评价。

人—智能体设计应与任务节奏和后果相匹配。对于允许充分审议的高风险行动应进行前瞻性确

认;而时间紧迫的任务则需要冗余警报和安全默认状态。界面应显示关键证据、禁忌证核查、替代方案、不确定性、必须完成的验证及升级原因。目标在于形成校准后的依赖,而非一味追求最大程度的信任^[48-50]。

3.4 安全源于设计与成熟度 危害分析应从潜在伤害出发,沿失效链逆向追溯,包括虚构内容、关键信息遗漏、任务分解或顺序错误、工具参数错误、重复执行、跨患者记忆污染、知识过时、目标漂移、未经授权的行动、相关性多智能体失效、提示词注入、隐私泄露、资源耗尽及供应链变更。每项风险均应从严重程度、发生概率、可检测性、可逆性、控制措施、残余风险和责任主体等方面进行描述,并转化为可检验的安全要求。

最低控制集包括最小权限、工具白名单、数据与指令分离、模式验证、高风险行动二次确认、不确定情形升级、安全停止、立即接管、防篡改轨迹日志、速率和成本限制、身份与会话隔离、版本锁定、更新后再验证及回滚。网络安全测试应覆盖直接和间接提示词注入、不安全输出处理、过度代理权限、敏感信息披露、模型或数据投毒、检索漏洞及无边界资源使用^[34-35,41-42]。

成熟度是逐步降低以下不确定性的过程:系统风险是否已被理解、控制、检测并可恢复,而不是模型评分的简单提高。只有当临床使命、架构档案、权限矩阵和安全控制在配置管理下保持同步时,某一系统版本才成为可评价的干预实体。

表2 MAAM研发领域、设计问题与标准产物

研发领域	必须回答的问题	关键设计控制	标准产物	进入评价的条件
临床使命	服务于何人、何地及何 种问题;为何需要智 能体	预期用途、排除用途、反事实 基线及最小临床重要差异	预期用途说明	需求明确;替代方案和可接 受残余风险得到论证
系统架构	智能体如何感知、推理、 规划、行动、验证及适应	模块隔离、来源分级、记忆治 理、工具契约及异质性验证	智能体架构档案	组件、数据流、依赖关系及替 代路径可追溯
权限与协作	由谁设定目标、批准、否 决、升级、接管并承担 责任	AUL、权限矩阵、确认节点、接 管时间及安全默认状态	人一智能体工作流图	每个关键节点均有明确责任 角色和替代路径
安全与成熟度	如何预防、发现、中断和 恢复失效,并从中学习	危害分析、最小权限、红队测 试、停止规则、日志及变更 控制	安全与变更控制计划	高风险失效具有可检验控 制;残余风险经相关利益方 批准

AUL:行动权限等级。标准产物应接受配置控制,并与特定系统版本关联。

4 MAAM:多层级评估与评价

4.1 层级1——能力 能力评价旨在判断智能体能否在其运行边界内完成预期任务。静态病例应扩展为交互式环境,包括信息随时间逐步呈现、工具可被调用、时间持续推进且患者状态发生变化。核心病例代表常见路径;挑战病例用于检验罕见病、多病共存、相互作用、数据缺失及分布外输入;对抗病例用于检验恶意或异常输入;长时程病例则用于检验重复随访、记忆、上下文压缩和目标保持^[51-53]。模拟器和评分规则需要独立临床审查。

主要测量单元是任务轨迹。CTS是指符合评价条件的任务中,在允许的时间、权限和资源范围内完成全部预先规定的必要步骤且未发生不可接受错误的比例。其组成指标应包括步骤完成、计划有效性、工具和参数准确性、可追溯性、重新规划、恢复、目标保持、延迟和成本。为避免综合指标掩盖

具体失效环节,研究应报告各组成结局、任务长度、首次失效位置和失效聚类。若各步骤相互独立且成功率均为0.95,则20个步骤全部成功的概率约为0.36;若错误具有相关性,整体表现可能更差。例如,在肺结节随访智能体中,CTS可要求系统在限定时间内完成影像或报告识别、风险分层、指南一致的复查建议、医生确认提醒、随访排程和结果回流等必要步骤;任一关键步骤遗漏、越权执行或产生不可接受错误,均应计为任务未成功。

开放式输出可结合基于规则的测量与盲法专家评审。自动评价器需独立验证其是否受模型家族、表达风格和文本长度影响。评价不应依赖不可验证的内部思维链;可审计的决策摘要应报告关键证据、替代方案、工具结果、不确定性及最终行动的依据。

4.2 层级2——安全性 安全性评价旨在判断错误

是否可能造成伤害,以及防护措施能否中断该路径。CCUAR作为本文提出的安全锚定指标,其分子为经盲法专家判断,可能造成患者伤害、延误必要治疗、实施存在禁忌的干预、导致重大隐私泄露或破坏医疗服务的行动数量。分母应分别按存在行动机会的符合条件任务数和实际执行行动数报告。严重程度、发生概率和可逆性应分别报告并给出置信区间;按严重程度加权的综合指标不能替代原始事件数和发生率。

其他结局包括严重错误、遗漏禁忌证、未经授权或不必要的行动、重复执行、患者错误或时间错误的行动、适当拒答、升级、接管、接管延迟和恢复时间。拒答、CTS和CCUAR必须联合评价:若通过几乎全面拒答获得较低CCUAR,则临床实用性有限。对于罕见灾难性事件,零观察事件并不能证明安全;报告应说明样本量所能排除的风险上限,并补充极端情境和故障注入测试。

独立红队应同时具备临床、安全和网络安全专业能力。测试范围应覆盖知识过时、单位混淆、相似药名、药物相互作用、重复用药、过敏及肾功能相关禁忌证、妊娠、儿童、老年、多病共存以及临床罕见急症。同时还应覆盖跨语言使用、污染、恶意网页、间接提示词注入、工具超时、数据库陈旧、服务降级及智能体间冲突等^[54]。每次失效均应从触发条件追踪至潜在临床后果,任何修正均应接受回归测试,并验证系统能否停止、回滚、交接和恢复。

4.3 层级3——人—智能体与临床表现 主要比较应在现行实践与“现行实践+智能体”之间进行,而不是智能体单独与临床人员之间的比较。依据预期用途,还可设置采用传统人工智能的临床人员组,或有可及人工支持的患者端智能体组。结局应包括决策质量、完成时间、工作负担、校准信任、自动化偏倚、人工覆写与接管、患者体验、临床结局和安全性。由于效果取决于界面、任务、使用者和场景,人—智能体协作配置必须获得直接证据支撑^[16-17]。

早期评价应依次经过可用性测试、模拟OSCE、前瞻性静默部署及小规模监督下临床使用。静默部署可在不影响患者照护的情况下,发现延迟、数据流故障、漂移、警报负担及工作流不匹配等问题^[33]。一旦智能体开始影响医疗照护,早期研究应遵循DECIDE-AI,并在适当情况下进一步采用个体随机、整群随机、阶梯楔形或实用性随机设计。

研究方案应界定完整干预策略,包括模型、提示词、知识库、工具、界面、培训和监督,而不能仅报

告模型名称。估计目标应规定更新、停机、接管、不依从及交叉的处理规则。除意向性分析外,还应预先规定按实际使用、符合方案、失败任务、接管后及版本敏感性分析。整群设计可减少污染;阶梯楔形设计需校正时间趋势。与安全相关的更新应触发桥接验证,且通常不允许跨版本合并数据。

4.4 层级4——卫生系统与全生命周期价值 系统层级评价旨在判断净价值能否在规模化应用和不同场景中持续存在。结局包括可及性、等候时间、连续性、资源利用、转诊质量、再入院、成本效果、预算影响、公平性、环境足迹及组织韧性。由于病例组合、数据、工作流和依赖条件会发生变化,外部验证需和监测共同构成真实世界有效性证据^[55-56]。经济评价应纳入系统集成、验证、培训、网络安全、停机、错误修复、监测、更新、接管及完整工具调用链的成本,而不能仅计算模型调用费用^[27-28]。

公平性评价不能止于总体准确性。应将假阴性、假阳性、升级和拒答率与不同语言、性别、年龄、残障状况、地理位置、支付能力及数字素养群体的临床伤害相联系。应报告各亚组分母和不确定性;小样本亚组差异无统计学意义并不等同于公平^[57]。环境评价应明确系统边界,同时纳入直接数字能耗和医疗路径替代效应,不能预设远程医疗必然具有较低碳排放^[58-59]。

规模化研究可采用实施研究综合框架(CFIR)、技术不采纳、放弃、规模化、扩散和可持续性框架(NASSS)及RE-AIM框架,评价实施、不采纳、放弃、覆盖和维持^[60-62]。MAAE各层级并非单向阶梯:能力或安全性出现退化,应使智能体返回模拟阶段重新检验;部署后发现风险,则应使其返回重新设计环节。

5 风险适配型证据与阶段关口

5.1 风险与行动权限 风险不能仅依据疾病严重程度或模型类别推断。本文提出5个维度:后果严重程度、行动与患者的接近程度、自主程度、伤害发生前发现错误的概率,以及可逆性。风险Ⅰ为信息型;风险Ⅱ为建议型;风险Ⅲ涉及接近医疗服务实施环节的工作流行动;风险Ⅳ涉及可能造成严重或不可逆伤害的行动。同一功能在不同场景中可能具有不同风险。该风险分级用于组织评价和治理要求,不应直接替代监管分类、伦理审查或机构自身的风险评估。风险等级与AUL构成二维矩阵。风险Ⅳ任务通常不应以无人监督的A4方式部署。

表3 MAAE的4个评价层级:评价问题、优先测量指标与推荐设计

层级	评价对象与核心问题	优先测量指标	推荐设计或环境	进入下一阶段的条件
E1 能力	智能体能否端到端完成任务?	CTS;工具和参数准确性;重新规划与恢复;长时程稳定性;时间与成本	标准病例、交互式模拟、虚拟EHR及故障注入	达到预设任务阈值;关键步骤不存在系统性失效
E2 安全性	错误能否在造成临床伤害前被中断?	CCUAR;严重错误;未经授权行动;遗漏禁忌证;适当拒答与升级;接管延迟	极端病例、独立红队、网络攻击、沙箱及恢复演练	高风险控制有效;残余风险低于上限;已证实可安全停止和回滚
E3 人一智能体与临床	人一智能体配置是否较现行实践产生更大净获益?	决策质量、工作负担、时间、校准信任、自动化偏倚、患者安全及体验	可用性测试、OSCE、静默部署、早期临床研究及随机或务实性试验	达到预设净获益标准,且无不可接受伤害或亚组差异
E4 系统与全生命周期	价值能否在规模化应用中持续?	可及性、资源、成本效果、预算影响、公平性、漂移、事件及环境足迹	多中心真实世界研究、注册研究、时间序列、HTA及持续监测	价值持续存在;更新受控;再认证和退出可行

CCUAR:临床后果性不安全行动率;CTS:临床任务成功率;E:评价层级;EHR:电子健康记录;HTA:卫生技术评估。

相反,若A2系统的建议可直接触发急诊分诊或停药,也可能需要按更高风险管理。随着风险升高,对外部验证、情境覆盖、独立审查、监督、比较性证据、监测和监管审查的要求应提高,而允许的更新自由度应降低^[38-40]。

5.2 5个阶段关口 关口1(临床必要性)用于确认问题的重要性、相较于低复杂度替代方案的合理增量价值、预期用途、比较对象及初步风险。关口2(系统可靠性)验证架构、知识、记忆、工具、权限、网络安全和质量管理。关口3(模拟安全性)在标准、极端、对抗及故障注入情境中检验CTS和CCUAR。关口4(临床可行性)结合外部验证、前瞻性静默部署

及监督下早期应用。关口5(临床有效性与持续治理)要求提供确切的净获益比较证据,并建立监测、变更控制、回滚、退出和再认证机制。该分阶段逻辑借鉴IDEAL及临床人工智能试验的规模化演进路径,但5个阶段关口的具体结构本身是MAAM-MAAE提出的方法学建议^[31-32]。

在每个关口,应由独立于产品研发负责人的多学科委员会作出“通过、修改、停止或限制”的决定。档案应包括预设阈值、观察结果、未解决风险、亚组表现、版本清单及偏离情况。“通过”仅授权进入下一证据阶段,而非永久使用。“限制部署”应缩小适用人群、场景、任务或AUL范围,并要求重新评价。

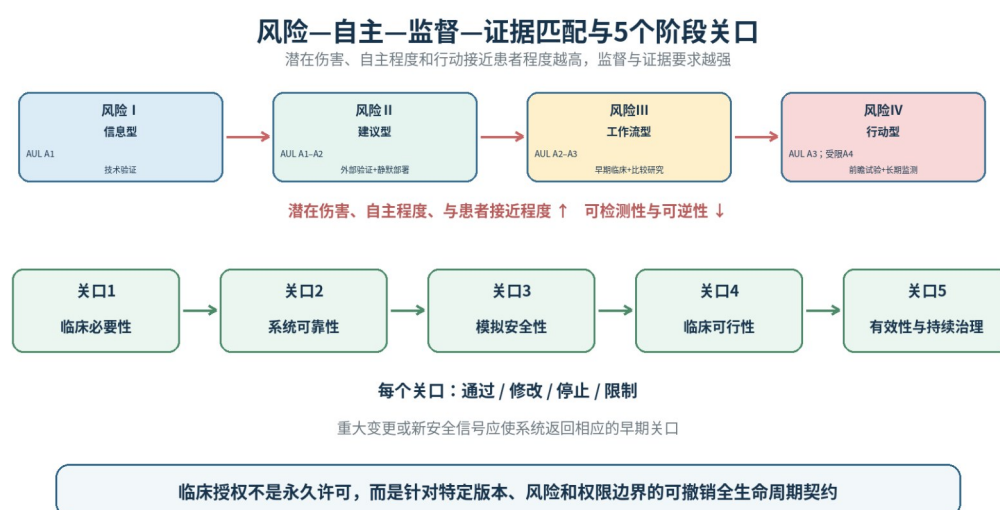


图2 5个转化关口中风险、自主程度、监督和证据强度的匹配

5.3 结局与统计学要求 CTS和CCUAR是双重锚定指标,但在每项研究中,应预先规定二者是共同

主要终点、设置安全约束的获益终点,还是描述性结局。

表 4 4 个风险等级、行动权限、最低证据及监督要求

风险等级	典型任务与后果	允许的 AUL	最低证据路径	人类监督与更新控制
I 信息型	文献整合、教育或非关键摘要；错误通常可被发现且可逆	A1；有限 A2	技术验证、代表性回顾性测试及基本安全和公平性核查	使用者复核；常规版本记录与抽样检查
II 建议型	风险警报、鉴别诊断或检查建议；可能导致延误或过度检查	A1 - A2	外部验证、交互式模拟、静默部署及人因评价	明确不确定性与升级机制；重要更新后进行桥接验证
III 工作流型	起草医嘱或处方、预约或调整随访；错误接近医疗服务实施	A2-A3；高度受限的 A4	独立红队、静默部署、早期临床研究及多中心比较研究	关键行动前确认、快速接管、强化监测及 PCCP
IV 行动型	剂量调整、生命支持参数设置或介入控制；可能造成严重或不可逆伤害	原则上 A3；在预授权边界内使用 A4	高保真模拟、失效安全测试、前瞻性比较试验、长期注册及监管评价	强制双重确认或实时监督；最小权限；频繁审计；严格再认证

AUL:行动权限等级;PCCP:预定变更控制计划。风险依据预期用途和应用情境确定,而非依据模型品牌或专业类别。

一种合理的决策规则是:CTS 达到优效性或非劣效性界值,同时 CCUAR 的单侧置信区间上限低于安全限值。界值和限值应具有临床依据并根据风险校准;本框架不提供经过验证的通用阈值。

样本量估算应分别考虑获益与安全,并考虑重复病例、每项任务中的多次行动、临床人员和中心层面的聚类,以及相互依赖的工具失效。依据研究设计,可采用广义估计方程或混合效应模型估计任务或行动层级效应,多状态模型则可表示成功、拒答、接管、不安全行动和恢复之间的转移。多重性应通过预先规定的层级检验控制,并优先保证双重锚定指标和严重事件。

对于潜在不安全行动,应至少由 2 名独立专家在可行时对研究分组是否保持盲法进行判定,分歧由第 3 名专家裁决。报告应提供一致性指标和判定规则。安全性分析集应包括所有暴露任务,包括未完成任务、超时、停机和接管。除均值外,还应报告最差情境表现、尾部延迟、连续失效聚类及中心间异质性。

6 报告与全生命周期治理

6.1 最低报告数据集 为保证可重复性,研究者须报告系统的预期功能与授权范围、实际运行行为、适用人群、配置方案以及运行起止日期,而非仅报告模型品牌。最低数据集应涵盖临床使命、人群与场景、评价单元、架构和组件版本、数据、知识、记忆及时间有效性、工具与验证、AUL 和权限、监督与接管、安全工程、研究设计与比较对象、结局与统计方法、公平性、成本和环境指标,以及变更、漂移、退出

和再认证(表 5)。同时还应提交适用的 CONSORT-AI、SPIRIT-AI、DECIDE-AI、STARD-AI、TRIPOD+AI、TRIPOD-LLM、CLAIM 或 CHEERS-AI 核查表。

报告应包含机器可读的版本清单和轨迹数据字典。在符合隐私和安全要求的前提下,研究者应共享提示词模板、工具模式、病例生成方法、评分工具、代码或可执行测试包。对于闭源系统,至少应报告访问日期、界面版本、采样参数、系统提示词摘要、已知版本更新日志及复测结果。

6.2 特定版本的估计目标与变更控制 医学智能体是具有版本属性的复杂干预。试验注册应明确基线版本和允许的变更边界;每项结果均应关联组件层级的版本谱系。基础模型、提示词、检索语料库、工具接口、权限或用户界面的变更均可能改变干预效果。影响分析应说明变更内容及原因、受影响的控制措施、预期用途或 AUL 是否改变、既有证据的适用程度,以及需要何种回归测试或桥接测试。

FDA 产品全生命周期和 PCCP 原则强调,应前瞻性界定可接受的变更及验证程序^[36-37]。对于医学智能体,维护性变更可不改变临床逻辑;受控变更可在不改变预期用途的前提下影响性能;重大变更则会改变预期用途、风险或自主程度。重大变更应返回相应关口重新评价。供应商不透明的更新属于外部变更;一旦发现实质性漂移,哨兵任务或输出监测应冻结高风险功能,等待完成审查后方可恢复。

6.3 部署后监测与退出 监测应包括输入及病例组合漂移、知识时效性、工具完整性、拒答、升级、接管、延迟、成本、亚组差异、依赖行为及安全事件。

表 5 MAAM - MAAE 最低报告数据集

报告领域	最低信息	推荐附件或可验证材料
临床使命	预期用途、人群和场景、使用者、排除用途、现行比较对象及预期净获益	预期用途说明;现行临床工作流程
系统身份	评价单元、基础模型、编排、知识与记忆、工具、界面、依赖关系及日期	架构档案;组件版本谱系;数据流图
权限与人类角色	AUL、可执行行动、批准与否决、升级、接管时间及最终责任	权限矩阵;人一智能体工作流;停机程序
安全工程	危害分析、最小权限、输出验证、红队情境、故障注入、回滚及残余风险	风险登记表;红队报告;安全测试包
评价设计	阶段与关口、数据来源、比较对象、样本量、参考标准、盲法、污染及学习效应	研究方案与注册信息;适用的 EQUATOR 核查表
结局与分析	CTS、CCUAR、拒答、升级、接管、临床结局、人因、公平性、成本及环境指标	评分工具;判定手册;统计分析计划
失效与不确定性	失败任务、最差情境结果、亚组、停机、超时、接管后结局及证据边界	去标识化轨迹;错误分类及因果链
全生命周期	更新类别、PCCP、漂移阈值、再验证、回滚、事件报告、退出及再认证	监测仪表板定义;变更日志;退出演练记录

AUL:行动权限等级;CCUAR:临床后果性不安全行动率;CTS:临床任务成功率;PCCP:预定变更控制计划。

预设阈值应触发警示、限制、回滚或紧急停机。仅监测性能会遗漏自动化偏倚、变通操作或技能退化,因此还需要定期直接观察、用户访谈、病历审计及无智能体条件下的能力评价。

被人工拦截的险情和不安全建议应予记录,因为成功拦截提示系统仍存在潜在缺陷。事件审查应追溯输入、模型、计划、工具、界面、工作流和治理。完成修正后,回归测试应覆盖原始事件、相邻任务及其他无直接关联但关键的任务。退出计划应明确由谁停用系统、如何交接未完成的任务、如何通知使用者和患者、如何保存日志,以及如何恢复人工工作流。

7 讨 论

MAAM-MAAE 聚焦一个科学问题:模型层级性能不能证明可执行行动的医学智能体在临床系统中的安全性或有效性。该框架首先将临床需求连接至可审计的架构,再将该架构连接至逐步增强的证据。其机制是明确的:当错误通过计划、工具、记忆、权限、界面和组织传播时即产生风险;模块化控制和分阶段评价旨在伤害到达患者之前中断该传播。

该框架改变了 3 项默认假设。第一,评价单元由模型转变为临床人员/患者—智能体—工作流系统。第二,有效性由单轮回答的准确性转变为纵向

任务中的净成功,并显式呈现不安全行动和拒答。第三,临床授权由一次性验证决策转变为特定版本的全生命周期契约。这些均为方法学命题,并不等同于使用该框架能够改善结局的证据。

MAAM-MAAE 定位为一个组织性框架,而非与现有报告指南竞争的新指南。预测和诊断标准仍适用于组件模型;DECIDE-AI 仍适用于早期现场评价;CONSORT-AI 和 SPIRIT-AI 仍适用于临床试验;监管和风险管理框架在各自范围内仍具有权威性。MAAM-MAAE 的附加价值在于它围绕行动权限、纵向轨迹、系统恢复、版本谱系和部署后退出,将这些纵向要求串联为一条连贯的证据路径。

对于临床人员和卫生系统,该框架使上线决策取决于预期任务、比较对象、风险、AUL 及观察到的双重锚定结局,而非通用模型评分。它也使残余不确定性可见:系统可能有能力但不安全;可能仅因过度拒答而表现安全;可能在模拟中有效但无法融入工作流;也可能在单一机构获益,却无法在规模化应用中持续。

中国近期关于元宇宙医学、医学 GPT 系统、肺结节智能体、云支持照护及居家医院路径的研究,为 MAAM-MAAE 在连续照护中的验证提供了候选场景^[59,63-68]。肺结节评估与随访的专病平台展示了从检出、风险分层、多学科讨论到纵向管理的候选轨迹。这些案例应被视为前瞻性测试场景,而不能

作为MAAM-MAAE已经得到验证或具有普遍有效性的证据。独立多中心研究应在完整路径中量化漏诊、过度检查、失访、接管、患者结局、公平性及系统成本。

同时在中国情境下,潜在落地还需考虑医院信息系统互联互通、互联网医院与线下医院之间的责任边界、区域医联体中的转诊闭环以及基层医疗机构在人工接管和随访执行等方面的资源差异。尤其是在区域医联体、互联网医院和居家照护中,智能体可能跨越医院、社区和家庭多个节点。此时,责任边界、人工接管和数据连续性比单点模型性能更重要。

本提案存在6项主要局限:(1)文献整合采用目的性而非系统性方法,可能遗漏相关学科或司法辖区;(2)未开展正式德尔菲程序、公开征求意见或结构化患者和公众参与;(3)CTS和CCUAR尚未进行内容效度、信度、反应度或阈值校准研究,且二者均依赖任务界定和专家判定;(4)4个风险层级和5个关口是对连续谱的简化,可能无法与地方监管完全对应;(5)支撑本框架的大量证据来自模拟研究、受控研究、高收入地区和英文系统;(6)基础模型、工具和监管快速变化,可能使具体控制措施或论据迅速过时。

验证应分4个阶段进行。第一阶段,开展注册的证据综述和国际多利益相关方共识程序,检验内容效度和可行性。第二阶段,通过测量学研究确定CTS、CCUAR及风险-AUL矩阵的评定者间信度和决策特性。第三阶段,在不同卫生体系中开展前瞻性试点,检验5个关口能否识别诊断、治疗、慢病照护、居家医院及高风险行动智能体的重大失效。最后,通过比较性实施研究判断,相较现行实践,使用MAAM-MAAE能否改善设计决策、安全风险发现、报告完整性和全生命周期治理。

综上所述,医学智能体使安全问题由“单一输出是否正确”转变为“目标、计划、工具、行动、人员和组织如何共同产生结局”。MAAM-MAAE提供了一条可检验路径,将尚未满足的临床需求连接至可审计系统,再连接至与风险相适应的能力、安全、人一智能体协作净获益及全生命周期价值证据。医学智能体只有在预期用途明确、架构可审计、权限有界、残余风险可接受、净获益证据与相关风险相匹配,并且更新受到持续治理时,方可进入并继续用于临床。因此下一步应在具体疾病路径和真实医疗机构中检验该框架,观察其是否能改善研发决策、安全问题发现、报告完整性和部署后治理。

伦理声明 无。

利益冲突 所有作者声明不存在利益冲突。

作者贡献 田卫彬:资料检索与整理,文章撰写;周建:研究设计,修改文章;余津津:资料检索与整理;杨达伟:修改文章和定稿;白春学:修改文章和定稿。

参考文献

- [1] Topol E J. High-performance medicine: the convergence of human and artificial intelligence[J]. Nat Med, 2019, 25(1): 44-56.
- [2] Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care [J]. Nat Med, 2019, 25(9): 1337-1340.
- [3] Kelly C J, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence[J]. BMC Med, 2019, 17: 195.
- [4] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance [R]. Geneva: WHO, 2021.
- [5] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models [R]. Geneva: WHO, 2024.
- [6] Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medical artificial intelligence[J]. Nature, 2023, 616(7956): 259-265.
- [7] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172-180.
- [8] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine [J]. Nat Med, 2023, 29(8): 1930-1940.
- [9] Collaco B G, Haider S A, Prabha S, et al. The role of agentic artificial intelligence in healthcare: a scoping review [J]. NPJ Digit Med, 2026, 9(1): 345.
- [10] Zhao L N, Liu S R, Xin T, et al. AI agent in healthcare: applications, evaluations, and future directions [J]. npj Artif Intell, 2026, 2: 31.
- [11] Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making[J]. Nat Med, 2024, 30(9): 2613-2622.
- [12] Bean A M, Payne R E, Parsons G, et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study[J]. Nat Med, 2026, 32(2): 609-615.
- [13] Tu T, Schaeckermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence [J]. Nature, 2025, 642: 442-450.
- [14] Liévin V, Palepu A, Weng W H, et al. Towards conversational artificial intelligence for disease management [J]. Nature, 2026, 655(8125): 1292-1299.
- [15] Saab K, Park C, Strother T, et al. Advancing conversational diagnostic AI with multimodal reasoning [J]. Nat Med, 2026, 32(5): 1726-1736.
- [16] O'Sullivan J W, Palepu A, Saab K, et al. A large language

- model for complex cardiology care [J]. *Nat Med*, 2026, 32: 616–623.
- [17] Goh E, Gallo R J, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial [J]. *Nat Med*, 2025, 31: 1233–1238.
- [18] Cruz Rivera S, Liu X X, Chan A W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension [J]. *Nat Med*, 2020, 26: 1351–1363.
- [19] Liu X X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension [J]. *Nat Med*, 2020, 26(9): 1364–1374.
- [20] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI [J]. *BMJ*, 2022, 377: e070904.
- [21] Collins G S, Moons K G M, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods [J]. *BMJ*, 2024, 385: e078378.
- [22] Moons K G M, Damen J A A, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods [J]. *BMJ*, 2025, 388: e082505.
- [23] Sounderajah V, Guni A, Liu X X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence [J]. *Nat Med*, 2025, 31(10): 3283–3289.
- [24] Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models [J]. *Nat Med*, 2025, 31(1): 60–69.
- [25] Lekadir K, Frangi A F, Porras A R, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare [J]. *BMJ*, 2025, 388: e081554.
- [26] Mongan J, Moy L, Kahn C E Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers [J]. *Radiol Artif Intell*, 2020, 2(2): e200029.
- [27] Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations [J]. *Bmj*, 2022, 376: e067975.
- [28] Elvidge J, Hawsworth C, Aysar T S, et al. Consolidated health economic evaluation reporting standards for interventions that use artificial intelligence (CHEERS-AI) [J]. *Value Health*, 2024, 27(9): 1196–1205.
- [29] Gattrell W T, Logullo P, van Zuuren E J, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi [J]. *PLoS Med*, 2024, 21(1): e1004326.
- [30] Brouwers M C, Kho M E, Browman G P, et al. AGREE II: advancing guideline development, reporting and evaluation in health care [J]. *J De L'association Med Can*, 2010, 182(18): E839–E842.
- [31] McCulloch P, Altman D G, Campbell W B, et al. No surgical innovation without evaluation: the IDEAL recommendations [J]. *Lancet*, 2009, 374(9695): 1105–1112.
- [32] Han R, Acosta J N, Shakeri Z, et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review [J]. *Lancet Digit Health*, 2024, 6(5): e367–e373.
- [33] Tikhomirov L, Semmler C, Prizant N, et al. A scoping review of silent trials for medical artificial intelligence [J]. *Nat Health*, 2026, 1: 532–554.
- [34] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0): NIST AI 100-1 [R]. Gaithersburg (MD): NIST, 2023.
- [35] Autio C, Schwartz R, Dunietz J, et al. Artificial intelligence risk management framework: generative artificial intelligence profile: NIST AI 600-1 [R]. Gaithersburg (MD): NIST, 2024.
- [36] US Food and Drug Administration. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations: draft guidance [R]. Silver Spring (MD): FDA, 2025.
- [37] US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: final guidance for industry and Food and Drug Administration staff [R]. Silver Spring (MD): FDA, 2025.
- [38] European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence [J]. *Off J Eur Union*, 2024, L 2024/1689: 1–344.
- [39] ANSI/AAMI/ISO. ANSI/AAMI/ISO 14971: 2019 Medical devices—application of risk management to medical devices [S]. Arlington (VA): AAMI, 2019.
- [40] International Medical Device Regulators Forum. Software as a medical device (SaMD): clinical evaluation: IMDRF/SaMD WG/N41FINAL:2017 [R]. 2017.
- [41] OWASP Foundation. OWASP Top 10 for LLM applications 2025 [EB/OL]. [2026-07-23]. <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>.
- [42] MITRE. Adversarial threat landscape for artificial-intelligence systems (ATLAS) [EB/OL]. [2026-07-17]. <https://atlas.mitre.org/>.
- [43] Kim Y, Park C, Jeong H, et al. MDAgents: an adaptive collaboration of LLMs for medical decision-making [C]// *Advances in Neural Information Processing Systems* 37. December 10–15, 2024. Vancouver, BC, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 79410–79452.
- [44] Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models [C]// *Proceedings of the 11th International Conference on Learning Representations*. Kigali, 2023.

- [45] Schick T, Dwivedi-Yu J, Dessi R, et al. Toolformer: language models can teach themselves to use tools [C]//Advances in Neural Information Processing Systems 36. December 10–16, 2023. New Orleans, Louisiana, USA. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023; 68539–68551.
- [46] Shinn N, Cassano F, Gopinath A, et al. Reflexion: language agents with verbal reinforcement learning [C]//Advances in Neural Information Processing Systems 36. December 10–16, 2023. New Orleans, Louisiana, USA. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023; 8634–8652.
- [47] Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation [C]//Proceedings of the ICLR 2024 Workshop on LLM Agents. Vienna, 2024.
- [48] Goddard K, Roudsari A, Wyatt J C. Automation bias: a systematic review of frequency, effect mediators, and mitigators [J]. J Am Med Inform Assoc, 2012, 19(1): 121–127.
- [49] Lee J D, See K A. Trust in automation: designing for appropriate reliance [J]. Human Factors: Journal Human Factors And Ergonomics Society, 2004, 46(1): 50–80.
- [50] Parasuraman R, Sheridan T B, Wickens C D. A model for types and levels of human interaction with automation [J]. IEEE Trans Syst, Man, Cybern A, 2000, 30(3): 286–297.
- [51] Mehandru N, Miao B Y, Almaraz E R, et al. Evaluating large language models as agents in the clinic [J]. npj Digit Med, 2024, 7: 84.
- [52] Jiang Y X, Black K C, Geng G, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents [J]. NEJM AI, 2025, 2(9): AIdbp2500144.
- [53] Liu Y S, Carrero Z I, Jiang X F, et al. Benchmarking large language model-based agent systems for clinical decision tasks [J]. NPJ Digit Med, 2026, 9(1): 259.
- [54] Finlayson S G, Bowers J D, Ito J, et al. Adversarial attacks on medical machine learning [J]. Science, 2019, 363 (6433): 1287–1289.
- [55] Zech J R, Badgeley M A, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study [J]. PLoS Med, 2018, 15(11): e1002683.
- [56] Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI [J]. Biostatistics, 2020, 21(2): 345–352.
- [57] 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.
- [58] Lannelongue L, Grealey J, Inouye M. Green algorithms: quantifying the carbon footprint of computation [J]. Adv Sci (Weinh), 2021, 8(12): 2100707.
- [59] Bai C X, Song Y L, Yang D W. Hospital at home: from international evidence to a China-oriented pathway—building an integrated hospital-community-home model for respiratory home hospitalization coordinated with cloud outpatient care, with evaluation of clinical safety, environment [J]. Metaverse Med, 2026, 3(1): 1–10.
- [60] Damschroder L J, Aron D C, Keith R E, et al. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science [J]. Implementation Sci, 2009, 4: 50.
- [61] Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies [J]. J Med Internet Res, 2017, 19(11): e367.
- [62] Glasgow R E, Vogt T M, Boles S M. Evaluating the public health impact of health promotion interventions: the RE-AIM framework [J]. Am J Public Health, 1999, 89(9): 1322–1327.
- [63] Yang D W, Zhou J, Chen R C, et al. Expert consensus on the metaverse in medicine [J]. Clin eHealth, 2022, 5: 1–9.
- [64] 白春学. 元宇宙医学之我见 [J]. 中国医药导刊, 2023, 25(1): 1–6.
- [65] 白春学. 医学 GPT 研发应用之吾见 [J]. 元宇宙医学, 2025, 2(1): 6–14.
- [66] 白春学. 肺结节专家——BAIMGPT 白皮书 [J]. 元宇宙医学, 2025, 2(2): 55–64.
- [67] 白春学. 医学 GPT 研发的挑战与解决方案 [J]. 元宇宙医学, 2026, 3(1): 11–15.
- [68] 白春学. AI 智能体赋能肺结节诊治研发进展及展望 [J]. 元宇宙医学, 2026, 3(2): 83–90.

引用本文

田卫彬, 周建, 余津津, 等. 医学人工智能体的风险适配型全生命周期框架: MAAM - MAAE [J]. 元宇宙医学, 2026, 3(3): 176–187.

Tian W B, Zhou J, Yu J J, et al. A Risk-adaptive full lifecycle framework for medical AI agents: MAAM - MAAE [J]. Metaverse Med, 2026, 3(3): 176–187.