

DOI:10.61189/877948hydwr

· 主编寄语 ·

医学智能体驱动医学改革:构建智能体医学新科学

顾建英¹, 白春学^{2,3,4,5*}

1. 复旦大学附属中山医院整形外科, 上海 200032
2. 复旦大学附属中山医院呼吸与危重症医学科, 上海 200032
3. 复旦大学附属中山医院 AI+肺癌防治中心, 上海 200032
4. 上海呼吸物联网医学工程技术研究中心, 上海 200032
5. 上海市呼吸病研究所, 上海 200032

[摘要] 医学人工智能正从预测与内容生成迈向目标驱动、工具调用和受控执行,由此产生了超越模型性能评价的新的科学问题。本文提出“智能体医学科学”(agentic medicine science):研究医学智能体如何在临床目标、人类监督与制度边界内参与医疗行动,以及这些行动如何影响患者结局、医疗过程与卫生系统绩效。其基本研究单元是患者—专业人员—智能体—医疗体系(patient-professional-agent-health system, PPAHS)复合系统,其知识结构由临床行动、临床证据和责任治理三大支柱构成,其全球转化遵循标准、科学、安全和普惠(4S)议程。本文区分医学 GPT、自动化 workflows、医学智能体、多智能体系统和具身智能体,提出从技术验证、临床评价、结局研究到持续监测的分级证据路径。肺结节闭环管理、云门诊协同居家医院等中国实践提示,医学智能体的价值不在于复制医生,而在于将分散的知识、任务与责任重构为可验证、可约束、可追责的连续照护系统。

[关键词] 医学智能体;智能体医学;大语言模型;临床评价;人工智能治理;人机协作;卫生体系改革

[中图分类号] R-3 **[文献标志码]** A

Medical agents and the rise of agentic medicine: a new clinical science and a global agenda for accountable action

Gu Jianying¹, Bai Chunxue^{2,3,4,5*}

1. Department of Plastic and Reconstructive Surgery, Zhongshan Hospital, Fudan University, Shanghai 200032, China
2. Department of Respiratory and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China
3. AI+Lung Cancer Prevention and Treatment Center, Zhongshan Hospital, Fudan University, Shanghai 200032, China
4. Shanghai Respiratory Internet of Things Medical Engineering Technology Research Center, Shanghai 200032, China
5. Shanghai Institute of Respiratory Diseases, Shanghai 200032, China

[Abstract] Medical artificial intelligence is moving beyond prediction and content generation toward goal-directed reasoning, tool use, and controlled execution, giving rise to new scientific questions that transcend model performance evaluation. We propose agentic medicine science, a discipline that investigates how medical agents participate in clinical action within defined objectives, human oversight, and institutional boundaries, and how such participation affects patient outcomes, care processes, and health system performance. Its fundamental unit of analysis is the patient-professional-agent-health system (PPAHS) composite system. Its knowledge architecture comprises three pillars: clinical action, clinical evidence, and accountability and governance; global translation is guided by a 4S agenda of standards, science, safety, and solidarity. We distinguish medical GPTs, automated workflows, medical agents, multi-agent systems, and embodied agents, and propose a tiered evidence pathway extending from technical validation and clinical validation to clinical outcomes research and continuous post-deployment monitoring. Chinese practices, including closed-loop pulmonary nodule management and hospital-at-home care coordinated with cloud-based outpatient services, suggest that the value of

[收稿日期] 2026-07-29

[接受日期] 2026-09-13

[基金项目] 四大慢病重大专项(2023ZD0509600、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 (2023ZD0509600, 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] 顾建英,博士,主任医师、教授; E-mail: gu.jianying@zs-hospital.sh.cn

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

medical agents lies not in replicating physicians, but in reorganising fragmented knowledge, tasks, and responsibilities into a continuous care system that is verifiable, controllable, and accountable.

[Key Words] medical agents; agentic medicine; large language models; clinical evaluation; artificial intelligence governance; human - AI collaboration; health-system reform

医学人工智能(artificial intelligence, AI)先后扩展了临床的感知、推理、预测与内容生成能力,但临床工作的关键并非一次性输出,而是随信息变化持续编排任务、衔接部门、识别异常并完成照护任务。高绩效医学的实现取决于专业判断、计算能力与工作流的恰当耦合,而非单个模型性能领先^[1-4]。由此,医学人工智能发展到智能体体系的必要性显而易见。医学智能体以基础模型为认知内核,维持任务状态、制定计划,调用电子健康记录(electronic health record, EHR)以及检验、影像、指南或排班工具,并根据反馈调整后续行动^[5],具有自驱动、自进化和长上下文等特性,已有大量系统设计和应用探索。MIRA、MedAgentBench、MedHELM分别利用受控EHR沙箱、临床衍生任务和真实临床任务等方式对医学智能体的全流程能力、工具调用和临床真实表现进行了评价^[6-8]。这些研究表明,行动型AI及其真实任务表现已成为可检验对象,但尚不足以支持真实世界授权:沙箱无法充分再现数据缺失、组织摩擦、患者偏好和罕见伤害。

评价问题因而必须从“能否正确回答问题”升级为:系统能否在限定权限内安全完成任务,能否为患者和卫生系统带来净获益,以及行动后由谁解释、监测并负责。若仍以问答准确率评价能够写入病历、发起转诊或影响治疗的系统,证据强度将低于行动权限。为了描述这种以患者—专业人员—智能体—医疗体系为研究和评价对象的新问题,本文提出“智能体医学科学”,研究医学智能体如何在临床目标、人类监督和制度边界内参与行动,以及这些行动如何改变患者结局、专业实践与卫生体系。其目标不是使AI更像医生,而是使知识、任务、人机关系和组织责任的重组可被检验、约束与治理。

1 为什么需要智能体医学新科学

1.1 范式断裂:错误输出已演化为错误行动 预测模型通常在输出层接受评价,生成式AI又带来幻觉、错误引用和表达不一致,医学智能体则把风险延伸至行动层。错误的工具调用可导致漏随访、重复检查、危险药物组合、资源误配或升级延误。更危险的是复合失效:轻微数据漂移、错误记忆、接口

异常与自动化偏倚在多步骤链路中累积,形成单点测试难以发现的系统性伤害。自驱动、自进化和长上下文记忆等机制还会导致错误的积累与传递,本为纠错和优化而设计的迭代机制可能会导致更复杂的错误和问题模式,也更加隐蔽、难于追踪和发现。

因此,在智能体医学科学中,安全对象应由模型扩展为“模型—工具—人—工作流—组织”整体。同时,为澄清新的研究对象的内涵与边界,本文认为医学智能体至少须具备可审计的目标保持、情境感知、临床规划、受控工具调用、反馈纠错和适当弃权能力。缺少关键环节者可以是医学GPT或自动化程序,但不宜被冠以智能体之名;否则,监管对象、证据门槛与责任主体的边界都会变得模糊。

1.2 概念边界:从模型能力到临床能动性 医学智能体的最低条件应同时包括目标导向、跨步骤状态维持、情境性规划、外部工具使用、行动反馈和边界控制。“能生成计划”不等于“能执行计划”,“可自动运行”也不等于具备临床能动性。临床能动性还体现在通过长短记忆自动聚焦当前问题和利用背景信息,通过自驱动实现自动的工具选择与试错优化,通过自进化记忆试错形成的经验,这些新的特性都不是单纯的模型能力所能涵盖,映射到医学应用中,需要专业的Harness来承载和保障。另外,多智能体系统只有在角色分工、通信协议、冲突消解和共同审计均明确时才可能提高可靠性;同源模型的重复投票反而可能把共同偏差包装为共识。具身智能体连接机器人、可穿戴设备或家庭终端,物理后果更难逆转,证据与监督门槛应最高^[5,9]。这些需要更加高阶的安全与质量体系实现治理。

1.3 新研究对象:PPAHS复合系统 智能体医学的基本研究单元是患者—专业人员—智能体—医疗体系(patient - professional - agent - health system, PPAHS)。患者提供目标、偏好与生活情境;专业人员负责临床判断、行动授权、持续监督和必要救济;智能体整合信息、编排任务并受控执行;医疗体系提供数据、规则、资源和问责结构。任一环节的效能均依赖其余环节,故“AI vs 医生”的竞赛范式不能解释系统真实价值。

PPAHS也改变因果归因:获益可能来自模型、

界面、流程或团队分工;准确率提高也可能被警报疲劳、流程中断或责任模糊抵消。因此,评价必须结合临床医学、实施科学、人因工程、卫生经济学、伦理法学与系统工程。

2 智能体医学科学的三大支柱

2.1 临床行动:自主性必须与可逆性和临床接近度联动 临床行动(agency)应按具体任务授权,而非笼统地贴上“自主”标签。权限可依次分为只读检索、形成草案、提出建议、排队待签、条件性执行和自动执行,并同步标注可逆性、紧迫性、临床接近度及失效可发现性。自主性越高、行动越接近患者且后果越不可逆,所需的证据、监督和冗余就越强。

最小权限须在架构层落实:限定数据域、工具白名单、时间窗口和操作范围;高风险行动实行双通道确认或专业签署;工具调用完整记录输入、输出、版本、时间与责任主体标识。面对信息缺失、证据冲突或分布外病例,系统应弃权、求助或升级。多智能体协作宜采用异质证据、独立工具和明确仲裁;写入病历、开立医嘱或控制设备时,还须以确定性规则和独立安全层阻断不可逆错误。

2.2 临床证据:从模型准确率转向任务成功与患者净获益 证据强度必须与拟授予的行动权限匹配。医学问答、病例摘要和模拟患者适合用于机制探索,却不能直接支持真实世界的临床权限;随机试验也显示,仅向医生提供大语言模型并不必然改善诊断推理^[10]。评价对象应是“专业人员—智能体—医疗—工作流”系统。

建议以临床任务成功率(clinical task success rate, CTSR)和临床后果性不安全行动(clinically consequential unsafe action, CCUA)作为锚点结局。CTSR 要求任务在规定时间、权限和资源内完整完

成;CCUA 记录可能改变诊断、治疗、转诊、监测或患者行为,并可能造成临床伤害的不安全行动。两者须与漏执行、越权、适当弃权、人工接管、时间成本、患者体验、公平性、资源使用和临床结局共同报告;平均准确率既会稀释低频高损失事件,也可能把过度检查误作性能提升。

证据路径宜依次包括技术验证、任务模拟、交互沙箱、前瞻性静默研究、辅助性临床研究、对照结局研究和部署后持续监测。DECIDE-AI、CONSORT-AI 及 TRIPOD+AI 为相邻阶段的研究报告提供依据^[11-13];智能体研究还应报告目标、记忆、工具、权限、执行轨迹、回滚、版本及多智能体通信。任务、机构、版本或权限发生实质改变时,应返回相应阶段重新验证。

2.3 责任治理:行动可以分布,责任不能消失 责任治理(accountability)须明确谁设定目标、配置权限、批准行动、监测结局、处置错误、决定更新并承担责任。医疗机构负责本地选择、验证和监测;开发者负责预期用途、技术文档、风险控制与版本变更;专业人员的责任范围应限于其有权且有能力审查的判断。若审查时间、信息或退出权不足,“人在回路”只会成为责任转移的橡皮图章。

治理应覆盖全生命周期:上线前完成数据谱系、威胁模型、亚组评价和停机标准;上线时限定用途、工具与用户,并配置接管、告知和事件上报;上线后监测漂移、自动化偏倚、异常调用和更新影响。基础模型、检索库、提示、接口或规则的实质更新均可改变风险。

FUTURE-AI 与 WHO 框架强调公平、可追溯、稳健、可解释和人类自主^[14-16];对智能体而言,这些原则必须转化为可测试的控制点。监管也应从批准静态产品转向治理持续变化的行动系统^[9,16]。



图1 智能体医学科学的PPAHS研究对象、三支柱与4S全球护栏

患者告知强度则应与自主程度、临床接近度、附加风险和可选择性相称。

3 本专辑的使命:建立从概念到结局的完整知识链

本专辑不是应用案例的并列陈列,而是致力于构建完整知识链:建立模型、 workflow、智能体、多智能体与具身系统的可操作概念体系;形成数据—知识—模型—工具—工作流的可复现研发方法;以任务、团队、患者和卫生系统为递进评价单元;将权限、审计、更新和责任固化为系统属性。

肺结节管理贯穿风险分层、检查选择、多学科决策、长期随访与异常升级;居家医院连接医院、社区和家庭的监测、治疗、护理与快速转诊。此类闭环场景既可检验智能体能否将离散建议转化为连续照护,也可用于重构医学教育、科研和医院运营:学习者需掌握何时依赖、何时质疑,研究者需审计证据链,管理者需评价净价值而非演示性能。

必须警惕3种失衡:重能力展示而轻临床比较,重自动化效率而轻患者价值,重算法责任声明而轻组织问责。跨机构外部验证和亚组分析应成为常规,因为总体公平性指标并不能保证分布转移下的局部安全^[17]。在资源有限地区,还须评价基础设施、语言文化、工作负荷转移和机会成本,否则智能体可能复制优质中心的流程与偏差,而非缩小能力差距。

4 智能体医学的4S全球议程

4.1 标准(Standards):用统一标准构筑领域发展基石 标准首先应统一医学智能体的最低定义,继而统一预期用途、权限、风险、证据阶段和报告要求。研究至少应披露基础模型及版本、知识源、记忆机制、工具清单、权限边界、系统提示、执行轨迹、失败与弃权、人工监督方式、多智能体通信、成本和更新策略。互操作可依托 FHIR、标准术语与可移植接口,但技术互通不能替代语义一致和责任互认;通用基座也不能成为规避医疗安全要求的理由^[9]。

4.2 科学(Science):以真实结局替代排行榜 科学议程应从一次性排行榜转向多中心前瞻性研究、务实随机试验和真实世界比较。核心问题不是哪个模型得分最高,而是哪些任务、团队配置与监督强度能为哪些患者带来净获益。研究应预注册用途和权限,设定 CUA 停机界值,报告阴性结果与失败病例,并分析模型—机构—人群交互。经济评价须纳入基础设施、接口维护、人工复核、事件处置和更新再验证,避免将成本转移误判为效率提升。

4.3 安全(Safety):把安全写入架构、流程和组织 最低安全包应包括最小权限、工具白名单、数据隔离、提示注入防护、身份认证、适当弃权、异常升级、人工接管、可回滚执行、不可篡改审计、版本锁定、漂移监测和更新后再验证。高风险系统还应由独立安全控制器监督模型行动,并针对错误患者、过期信息、冲突医嘱、药物过敏、设备离线和资源不可用等情境开展对抗测试。事件分析必须同时追溯任务设计、界面、人力和组织激励,因为安全是 PPAHS 的涌现属性。

4.4 普惠(Solidarity):让智能体缩小而不是复制能力差距 普惠要求优先面向基层、资源不足地区、老年及多病共存人群和慢病连续管理。系统应具备本地语言支持、低带宽运行与离线降级能力,并支持区域知识适配、开放接口和共享安全基础设施,同时报告数字可及性、照护连续性与差距变化。公共价值不在于使少数中心再快几个百分点,而在于让高质量流程在保留地方自主和患者选择的前提下,做到可复制、可负担、可持续。

4.5 从中国实践走向可验证的全球贡献 中国的大规模临床场景、互联网医院、医疗物联网和分级诊疗实践,为系统创新提供了独特条件。白春学教授团队从医学 GPT“四改”策略和 BAIMGPT 肺结节专病实践^[18-19],推进至云门诊协同居家医院^[20]和肺结节智能体闭环管理^[21]。“数据清洗转向数据精选、简单问答转向专家分身、盲目信任转向质控核对、单纯循证转向循证与大医经验结合”,构成专病智能体知识治理与临床可用性的本土研发假设^[18-19]。

关键不在于扩张概念,而在于将这些假设置于可进行国际比较的证据框架中:以多中心外部验证检验迁移性,以静默和前瞻性研究验证真实数据流,以对照试验评价团队表现、患者结局和成本,并建立部署后登记系统,持续监测版本变更与安全事件。肺结节智能体可围绕未来肺癌风险,连接筛查、分层、多学科决策、随访、异常升级和家庭参与;居家医院智能体可检验医院—社区—家庭之间的任务连续性与快速转诊机制。只有将场景规模转化为透明方法、严谨证据和可共享标准,才能形成具有全球意义的学术增量。

综上所述,医学智能体标志着医学 AI 由生成知识转向参与行动,也迫使医学界重建概念边界、证据体系和责任制度。必须回答:何种系统才是医学智能体,哪些任务可被授权,何种证据足以支持何种权限,哪些权力须保留给人类,哪些责任不能被技术稀释。本专辑的使命不是宣告自主医疗时代

到来,而是建立一条可信路径:让智能体增强而非替代临床判断,连接而非割裂连续照护,扩大而非压缩患者自主,缩小而非复制卫生差距。只有把行动置于证据之内、自主置于边界之内、创新置于责任之内,医学智能体才可能成为改善健康的公共基础设施,而非未经验证的技术权力。

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 顾建英:素材整理、文章撰写;白春学:设计研究框架、文章审核。

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] Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medical artificial intelligence[J]. *Nature*, 2023, 616(7956): 259–265.
- [3] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine[J]. *N Engl J Med*, 2023, 388(13): 1233–1239.
- [4] Teo Z L, Thirunavukarasu A J, Elangovan K, et al. Generative artificial intelligence in medicine[J]. *Nat Med*, 2025, 31(10): 3270–3282.
- [5] Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare [J]. *Nat Biomed Eng*, 2025, 9(4): 432–438.
- [6] Ferber D, Hilgers L, Höper C, et al. Towards autonomous medical artificial intelligence agents [J]. *Nature*, 2026, 655(8125): 1282–1291.
- [7] Jiang Y, Black K C, Geng G, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents[J]. *NEJM AI*, 2025, 2(9): A1dbp2500144.
- [8] Bedi S, Cui H J, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM [J]. *Nat Med*, 2026, 32(3): 943–951.
- [9] Freyer O, Jayabalan S, Kather J N, et al. Overcoming regulatory barriers to the implementation of AI agents in healthcare[J]. *Nat Med*, 2025, 31: 3239–3243.
- [10] Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial[J]. *JAMA Netw Open*, 2024, 7(10): e2440969.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models [R]. Geneva: World Health Organization, 2025.
- [16] World Health Organization. Regulatory considerations on artificial intelligence for health [R]. Geneva: World Health Organization, 2023.
- [17] Yang Y Z, Zhang H R, Gichoya J W, et al. The limits of fair medical imaging AI in real-world generalization[J]. *Nat Med*, 2024, 30(10): 2838–2848.
- [18] 白春学. 医学 GPT 研发应用之吾见[J]. *元宇宙医学*, 2025, 2(1):6–14.
- [19] 白春学. 肺结节专家——BAIMGPT 白皮书[J]. *元宇宙医学*, 2025, 2(2):55–64.
- [20] 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.
- [21] 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

顾建英,白春学. 医学智能体驱动医学改革:构建智能体医学新科学[J]. *元宇宙医学*, 2026, 3(3):147–151.

Gu J Y, Bai C X. Medical agents and the rise of agentic medicine: a new clinical science and a global agenda for accountable action[J]. *Metaverse Med*, 2026, 3(3):147–151.