

DOI: 10.61189/066985xintdh

· 综述 ·

迈向人机共生医学：医学智能体时代的共生临床契约

杨达伟^{1,2,3,4}, 宋元林^{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 (artificial intelligence, AI) 正在由执行单一预测的算法, 演变为可持续感知情境、分解目标、调用工具、协同沟通并追踪结果的医学智能体。技术能力的跃迁使临床问题从“AI能否替代医生”转向“患者、医生、智能体与医疗机构如何共同形成判断、授权行动并承担后果”。现有研究显示, 大语言模型 (large language model, LLM) 在医学问答、鉴别诊断与管理推理中可达到较高水平, 但“模型优于人”并不等同于“人机组合优于人”: 随机试验和影像学研究同时揭示交互设计、自动化偏倚、错误放大、责任扩散与能力退化等系统风险。本文提出共生临床契约 (symbiotic clinical compact, SCC), 以共享智能、有界行动、患者共治和责任追溯4项契约, 规范患者—医生—智能体—医疗体系的权利、权限与责任; 提出界定问题 (Frame)—智能体分析 (Analyse)—共同审议 (Deliberate)—人类授权 (Authorise) 的 FADA 临床闭环, 以及从人工审核到人机共生医学的四阶段转化路径; 并以人机共生增益 (human - AI symbiotic gain, HASG) 和共生相关伤害 (symbiosis-related harm, SRH) 作为双锚点评价终点。SCC 不是赋予智能体无限医疗权力, 而是将任务自主性嵌入可撤回授权、风险分层、持续监测和机构问责。未来研究应从静态模型验证转向真实世界组合系统评价, 从平均准确率转向患者结局、分层公平、关系质量、专业能力与全生命周期安全。人机共生医学的目标, 是让智能体承担信息与流程复杂性, 让医生聚焦判断、责任与关怀, 让患者成为医疗决策的共同主体, 并使医疗体系由间断服务转向可学习、可追溯的连续健康伙伴。

[关键词] 医学智能体; 大语言模型; 人机协作; 共同决策; 患者自主; 临床治理; 责任追溯; 人工智能伦理

[中图分类号] R 4 **[文献标志码]** A

Toward human - AI symbiotic medicine: a symbiotic clinical compact for the era of medical agents

Yang Dawei^{1,2,3,4}, Song Yuanlin^{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 (AI) is evolving from algorithms that perform isolated predictive tasks into medical agents capable of continuously perceiving clinical context, decomposing goals, invoking tools, coordinating communication, and tracking outcomes. This transition shifts the central clinical question from whether AI can replace physicians to how patients, physicians, agents, and healthcare institutions can jointly form judgments, authorise actions, and assume responsibility for their consequences. Existing evidence indicates that large language models (LLMs) can perform strongly in medical question answering, differential diagnosis, and management reasoning. Yet “a model outperforms a human” does not mean that “a human - AI team outperforms a human alone”: randomised trials and imaging studies have exposed systemic risks arising from interaction design,

[收稿日期] 2026-07-20

[接受日期] 2026-09-09

[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] 杨达伟, 博士, 副主任医师; E-mail: yang.dawei@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

automation bias, error amplification, diffusion of responsibility, and erosion of professional competence. We propose the Symbiotic Clinical Compact (SCC), comprising four interrelated commitments—shared intelligence, bounded action, patient co-governance, and traceable accountability—to define the rights, authority, and responsibilities of patients, physicians, agents, and healthcare systems. We further introduce the FADA clinical loop—Frame, Analyse, Deliberate, and Authorise—together with a four-stage translational pathway extending from human review to human - AI symbiotic medicine. Human - AI symbiotic gain (HASG) and symbiosis-related harm (SRH) are proposed as dual-anchor evaluation endpoints. The SCC does not grant agents unrestricted authority to practise medicine. Instead, it embeds task-level autonomy within revocable authorisation, risk stratification, continuous monitoring, and institutional accountability. Future research should move beyond static model validation towards real-world evaluation of integrated human - AI systems, and beyond average accuracy towards patient outcomes, subgroup fairness, relational quality, professional competence, and lifecycle safety. The goal of human - AI symbiotic medicine is to enable agents to manage informational and procedural complexity, physicians to focus on judgment, responsibility, and compassionate care, patients to become genuine partners in medical decision making, and healthcare systems to evolve from episodic service providers into continuously learning, traceable health partners.

[Key Words] medical agents; large language models; human - AI collaboration; shared decision making; patient autonomy; clinical governance; traceable accountability; artificial intelligence ethics

医学从来不是纯粹的信息处理。诊断与治疗既依赖证据、概率和模式识别,也依赖对患者经历、价值取向、家庭资源、社会处境与不确定性的理解和解释。过去十年,深度学习在影像、病理和风险预测等特定任务上表现出接近专业人员的性能;但外部验证不足、报告质量参差和临床效用证据有限,使“高准确率”与“改善健康结局”之间始终存在转化鸿沟^[1-7]。生成式AI与LLM进一步改变了人机界面:系统不再只输出一个分数,还能通过自然语言解释、追问、拟定计划并调用外部工具^[8-15]。当系统可以自主选择下一步操作,访问电子健康记录(electronic health record, EHR),发送信息、生成医嘱草案或协调随访时,它已由知识工具迈向医疗过程的行动参与者。

这种变化使传统的“医生使用工具”模型不足以描述新型风险。算法错误可能经由可信的语言包装而更难被察觉;当多个模型、数据库、应用程序接口与人员共同作用时,单个输出与最终后果之间的责任边界可能变得模糊;患者既可能获得连续、个体化支持,也可能在不知情时被纳入自动化流程。世界卫生组织(WHO)已将自主、隐私、透明、公平、问责与持续评估置于医疗AI治理核心;欧盟《人工智能法案》和美国食品药品监督管理局(FDA)的全生命周期思路,也推动高风险AI从一次性准入转向变更控制与部署后监测^[16-20]。然而,现有规范多以模型、产品或研究报告为对象,尚缺少一个面向日常临床协作、能够同时约束患者、医生、智能体和机构的简洁契约。

本文据此提出SCC。其理论起点不是人机能力竞赛,而是社会技术系统设计:临床智能可以分布,

临床行动必须授权;任务可以协同,责任不能蒸发。SCC试图把人机共生从愿景转化为可实施的权利清单、权限结构、临床路径和评价框架。

1 从AI辅助到人机共生医学

1.1 医学智能体构成一次范式跃迁 传统临床AI通常在预定义输入与单一输出之间建立映射,例如识别影像异常或估计事件风险。LLM提供跨任务语言接口,而医学智能体在此基础上增加目标维持、记忆、规划、反思、工具调用和行动能力,可在多轮交互中根据新信息调整策略^[8-15,21]。因此,区分“模型”与“智能体”至关重要:模型回答问题,智能体围绕目标组织一系列状态依赖的操作;模型风险主要来自错误输出,智能体风险还包括错误目标、错误任务分解、错误工具调用、权限越界、记忆污染和级联失败。

医学智能体的临床价值不在于模仿一个全能医生,而在于连接原本碎片化的知识、数据与流程。它可以在门诊前整理病史,在诊疗中检索证据并比较方案,在诊后协调检查、用药教育、风险预警与家庭随访;也可在专病路径中调度影像模型、风险计算器、指南库和患者报告结局指标。白春学团队提出的元宇宙医学、医学GPT、肺结节智能体,以及云端门诊—医院—社区—家庭协同路径,体现了从知识问答走向专病闭环和连续照护的技术路线^[22-28]。然而,连接越广,系统的实际影响越接近医疗权力,治理对象也越必须从单一算法扩展为完整工作流。

1.2 性能领先不等于临床共生增益 LLM在医学知识与复杂病例上的进展令人瞩目。Med-PaLM系列在医学问答中持续提升^[10-11],AMIE在模拟文本问诊

和复杂鉴别诊断中显示出较强能力^[29-30]；其他研究也显示,生成式 AI 在患者问答与复杂诊断推理任务中可达到较高表现^[31-33]。然而,组合效果取决于人机如何相互影响。在 Goh 等开展的随机试验中,医生获得 LLM 辅助后,诊断推理得分并未显著提高,而 LLM 单独表现更高；另一项管理推理试验则观察到,LLM 辅助使医生得分提高 6.5 个百分点,但耗时也相应增加^[34-35]。一项纳入 140 名放射科医师的研究显示,AI 辅助效应高度异质,经验年限、亚专科和既往 AI 经验均不足以可靠预测受益者,错误的 AI 建议还会降低总体及多个任务的表现^[36]。乳腺影像研究^[37]进一步表明,不同资历的医师均可能受到错误建议所致自动化偏倚的影响。

这些结果否定了“把更强模型放到医生旁边即可自然增益”的线性假设。实验研究表明,人机组合可以优于任一成员单独工作,但这种互补增益依赖对任务、信息和分工的恰当设计^[38]。临床效果由模型性能、任务难度、信息呈现顺序、医生基线能力、时间压力、信任校准、患者偏好和机构流程共同

塑造。因此,共生医学关注的是交互产生的增量价值,而不是模型、医生与患者能力的简单相加。

1.3 评价单元必须扩展为四元系统 本文将人机共生医学定义为:以患者目标和权利为中心,由医生、患者、医学智能体与医疗机构共同参与信息整合、临床推理、决策实施和持续管理,并以可撤回授权、人工接管与全程追溯约束智能体行动的医疗模式。其最小评价单元是患者—医生—智能体—医疗体系 (patient - clinician - agent - health system, PCAHS),而非孤立模型。

PCAHS 视角揭示 4 个常被忽略的事实:(1)数据并非脱离患者而存在,患者对数据用途、自动化程度和人工复核具有正当利益;(2)医生不仅是 AI 输出的接收者,也是问题界定者、证据解释者和责任承担者;(3)智能体的“自主”是组织授予的任务权限,而不是独立的道德或法律主权;(4)机构决定采购、配置、培训、监测与事故处置,因此不能把系统性风险单独压给一线医生。

表 1 传统医学、AI 辅助医学、智能体医学与人机共生医学的范式比较

维度	传统医学	AI 辅助医学	智能体医学	人机共生医学
基本单元	医患关系	医生+单任务模型	用户+目标导向智能体	PCAHS 四元系统
AI 功能	无,或信息检索	识别、预测、提示	规划、记忆、工具调用、执行	在契约和权限内协同推理与行动
时间结构	就诊事件驱动	单次任务嵌入	多轮任务持续	跨医院—社区—家庭连续闭环
决策权	医生主导	医生采纳或拒绝建议	可能出现事实性自动执行	患者参与、医生授权、机构治理
主要风险	信息不足、认知偏差	泛化失败、误用、自动化偏倚	越权、级联错误、目标漂移	关系失衡、责任扩散、数字分层
证据标准	专业规范与临床结局	模型性能与外部验证	任务成功率与工具可靠性	HASG、SRH、患者权利与系统韧性
成熟标志	专业能力	稳定辅助	可完成多步任务	净获益可证、边界可控、责任可追

AI:人工智能;HASG:人机共生增益;PCAHS:患者—医生—智能体—医疗体系;SRH:共生相关伤害。

2 共生临床契约:从能力分工到权利、权限与责任

SCC 由 4 项相互制约的契约组成。共享智能回答“谁贡献何种知识”;有界行动回答“智能体能做什么”;患者共治回答“为谁行动、由谁定义目标”;责任追溯回答“出现后果时谁必须解释和纠正”。四者缺一不可:没有共享智能,AI 退化为静态工具;没有有界行动,协作可能转化为越权;没有患者共治,效率可能牺牲自主;没有责任追溯,多方协作会制造责任真空。这一结构回应了医疗 AI 实施中反复出现的伦理、责任与监管难题^[39-41]。

2.1 共享智能:以互补和可质疑性替代意见相加 医生贡献情境判断、因果解释、风险—获益权衡、异常识别和伦理判断。患者贡献症状体验、生活目

标、风险偏好、可承受成本与治疗意愿。智能体贡献跨源数据整合、证据检索、风险计算、遗漏提醒、方案模拟和连续监测。共享智能的机制不是多数投票,而是结构化互相质询:智能体必须呈现证据来源、关键假设、不确定性与替代解释;医生应检验输出是否适用于当前患者,并主动寻找反证;患者应能指出目标偏差、事实错误或不可接受的负担。

可质疑性比一般意义的“可解释性”更接近临床需要。冗长的自然语言解释可能只是流畅叙事,并不保证因果可靠。真正有用的界面应支持溯源至原始记录,区分观察事实与模型推断,显示信息缺失和置信边界,并允许医生和患者提出“如果……会怎样”的反事实问题。对于高风险建议,应通过独立证据通道或相互独立的模型进行核验,而

不是让同一模型自我确认。

2.2 有界行动:任务自主性不等于医疗主权 智能体权限应按风险、可逆性、可检测性和情境稳定性分层,而不能因技术能力提高就一次性开放。信息读取与整理通常风险较低,提出建议会影响认知锚定,生成医嘱草案已接近行动入口,自动执行则可能直接改变患者状态。权限应动态调整:当数据缺失、分布漂移、模型升级、患者病情恶化或监测信号异常时,系统应自动降级并转交人工处理。

高风险、不可逆或价值负载型事项原则上保留人类授权,包括最终诊断确认、处方或剂量改变、侵入性操作、生命支持、限制自由的措施、重大伦理选择及与患者目标冲突的方案。即使在低风险自动执行场景,也必须预设停止条件、操作上限、复核频率和回滚机制。这里的人在回路(human-in-the-loop, HITL)不是界面上的形式点击,而是具备时间、信息、能力和实际否决权的有效监督。

2.3 患者共治:把知情同意升级为持续授权 患者不应只是智能体服务对象,而应参与目标设定、数据授权、偏好表达、方案选择、输出复核、错误报告和疗效评价。传统一次性知情同意难以覆盖可更新模型和连续智能体,因此需要“持续授权”:明确智能体身份、用途、数据范围、自动化等级、第三方参与、模型更新及退出方式。当用途、权限或风险发生实质变化时,需要重新征求授权。

最低患者权利包括知情权、选择权、拒绝权、解

释请求权、人工复核权、数据纠错权和申诉权。患者拒绝使用智能体,不应因此遭受不合理的服务降级或惩罚。对于认知障碍、语言障碍、低数字素养或资源受限人群,机构应提供可理解材料、代理决策支持和线下替代路径。患者研究显示,安全、选择、偏倚、数据保护、责任和人际互动是主要关切;保留临床人员参与、明确监督责任并提供选择,有助于提升接受度^[42-44]。因此,患者共治既是伦理要求,也是建立校准信任的实施条件。

2.4 责任追溯:行动可以分布,责任不能消失 智能体没有独立承担专业责任的能力。医疗机构应对采购选择、预期用途、权限配置、系统集成、人员培训、运行监测、版本变更与事故响应承担组织责任。开发者应对设计、数据、验证、网络安全、已知限制和更新透明负责。医生对依赖专业判断并由其授权的行动负责。患者负责表达自身偏好并作出授权,但不应承担算法缺陷或机构治理失败的责任。相关伦理研究同样强调:即使 AI 提供第二意见,也必须保留可识别的人类责任与解决分歧的机制^[45]。

责任追溯需要“可重建的临床事实链”,至少记录任务目标、输入来源、模型与版本、提示和工具调用、候选方案、不确定性、人工修改、患者偏好、授权人、执行结果及异常处置。记录的目的是无限留存全部对话,而是在数据最小化前提下支持临床解释、质量改进和法律审查。

表 2 SCC 四项契约下 PCAHS 各方的核心权利、权限与责任

契约	患者	医生/临床团队	医学智能体	医疗机构/开发者
共享智能	提供目标、体验与偏好;质疑错误	界定问题、解释情境、核验证据	整合数据;给出来源、假设、不确定性与备选项	保证数据质量、互操作性与知识更新
有界行动	选择自动化等级;可随时退出	配置任务边界;在关键节点授权/接管	仅执行白名单任务;触发条件下停止或降级	按风险分层授权;设置回滚、停机和冗余
患者共治	知情、选择、拒绝、复核、纠错与申诉	促进共同决策;不得以 AI 替代沟通	使用可理解语言;尊重偏好与撤回	提供无惩罚替代路径和无障碍支持
责任追溯	获知责任主体与事故处理结果	记录判断、修改和授权依据	保存可审计操作链;标记版本与工具	明确系统负责人;监测、报告、赔偿并改进
不可转移底线	不承担系统缺陷责任	不得盲签或形式复核	不得自授权限或隐瞒不确定性	不得把系统性风险单压给一线医生或患者

SCC:共生临床契约;PCAHS:患者—医生—智能体—医疗体系。

3 进入临床:最小共生单元与 FADA 闭环

3.1 从最小共生单元起步 人机共生医学不应从“无人医院”起步,而应从边界清晰、价值可测的最小共生单元(minimum symbiotic unit, MSU)起步。

一个合格的 MSU 应包含以下要素:明确的临床目标和适用人群;一名可识别的责任医生或团队;经目标场景验证的智能体;允许与禁止任务清单;至少一个患者参与节点;人工接管与故障安全机制;预设的获益、伤害和公平指标。



图1 共生临床契约(SCC)与FADA临床闭环

4项契约界定权利、权限与责任;FADA把智能体输出与临床行动以共同审议和人类授权隔开,并由执行结果持续反馈。

MSU把宏大愿景压缩为可测试的临床微系统,使风险与收益能够在扩展前被观察和校正。

任务选择应优先考虑高频、规则较清晰、后果可逆、可被及时复核且当前存在明显流程负担的场景,如病历摘要、指南匹配、低风险随访、检查提醒和异常结果升级。对于诊断定论、处方自动调整或急危重症处置,应在获得充分前瞻性证据和监管许可前保持更高层次的人工控制。肺结节管理是典型的MSU候选:相关信息横跨影像、病理、危险因素和随访时间轴,适合由智能体整合与提醒;但恶性风险解释、侵入性检查和手术选择仍需专科团队与患者共同决策^[25,28]。

3.2 FADA:把技术输出与临床行动隔开 FADA闭环将临床智能分为4个不可跳过的阶段。(1)界定问题(Frame):由医生与患者共同明确临床问题、时间范围、价值目标和不可接受结局,避免系统优化错误目标。(2)智能体分析(Analyse):由智能体汇集记录、检索证据、计算风险、生成备选方案,并明确缺失信息与不确定性。(3)共同审议(Deliberate):由患者、医生与智能体共同权衡获益、伤害、可行性、成本和价值一致性;智能体在此承担“证据与反证生成器”,而非裁决者。(4)人类授权(Authorise):由具

备资质的专业人员批准、修改或拒绝行动,并指定监测指标和停止条件。

授权后不是闭环终点。执行结果进入监测,若出现新证据、病情改变、患者撤回授权或系统漂移,必须重新界定问题。FADA的关键是设置“语义防火墙”:智能体的建议不能因被写入EHR或医嘱界面就自动获得临床权威;只有经过共同审议和有效授权,技术输出才转化为行动。

3.3 四阶段转化与权限升级 第1阶段为人工逐项审核:智能体只完成局部、低风险任务,重点验证可用性和错误谱。第2阶段为人机协作:双方共同形成判断,但所有行动仍由医生执行。第3阶段为受监督智能体照护:智能体可在白名单内自动执行可逆、低风险任务,异常和高风险节点必须强制升级。第4阶段才是跨机构、跨场景的共生医学:信息和任务可在医院、基层与家庭之间连续流动,患者权限和责任链随流程迁移,高风险决策仍由人类承担。

权限升级不能以“试用期间没有事故”为唯一依据,而应同时满足外部验证、前瞻性真实世界评价、亚组公平、网络安全、人员胜任、患者接受、稳定监测和可回滚等条件。任何模型重大更新、用途扩展或工作流迁移,均应触发重新验证。FUTURE-AI

可支持可信设计与部署,DECIDE-AI 聚焦早期临床评价,SPIRIT-AI/CONSORT-AI 规范试验方案与报告,FDA 全生命周期框架则支持变更管理^[19-20,46-52];医学智能体还需额外报告目标、工具调用、权限边界和多步失败。

3.4 医生角色与组织能力重构 智能体时代,医生的核心价值将从“记住最多知识”转向提出正确问题、识别适用边界、整合冲突证据、管理不确定性、理解患者价值、监督系统行为和处理异常。医生必须同时具备无 AI 条件下的独立临床能力与有 AI 条

件下的协作能力;后者包括提示设计与问题表述、来源核验、自动化偏倚识别、反事实检查、权限管理和事件报告。若培训只教“如何更快使用 AI”,而不教“何时停用 AI”,将加速依赖与能力退化。

组织能力同样关键。机构需建立临床、信息、工程、伦理、法务、患者代表共同参与的治理结构,维护智能体登记册、用途和权限清单、版本记录、风险分级、性能仪表盘及事故响应流程。真实世界部署应视为持续质量改进与临床研究的结合,而非采购完成后的普通软件上线。

表 3 医学智能体四阶段临床转化路径及最低升级条件

阶段	允许行动	人类控制	核心证据	最低升级条件
I 人工审核	摘要、检索、风险提示、 草拟	逐项核对,不直接 执行	技术验证、外部验证、错误谱	预期用途清晰;无严重不可控 错误;用户可识别失败
II 人机协作	交互式推理、方案比较、 任务编排	医生形成并执行最 终决定	前瞻性小规模真实临床评价	临床效用与人因可接受;患者 告知;责任主体明确
III 受监督 照护	白名单内低风险、可逆任 务自动执行	抽查+异常强制接管	比较研究、亚组安全、运行稳 定性	阈值、停机、回滚、审计、网络安 全及补救机制齐备
IV 共生医 学	跨场景连续协调与个体 化管理	高风险节点授权;系 统级监督	患者结局、长期净获益、卫生经 济与公平证据	跨机构互认;持续监测;变更控 制;患者授权可携带

阶段代表治理成熟度而非技术先进性;不满足升级条件时应降级运行。

4 评价人机共生医学:双锚点、组合比较与全生命周期证据

4.1 人机共生增益:证明组合系统的真实增量价值 本文提出 HASG,指 PCAHS 相较于当前最佳常规照护所产生的、经风险与资源调整后的净增量价值。HASG 不能用模型准确率替代,而应覆盖患者结局、决策质量、照护连续性、患者参与、医生工作负担、可及性、公平和成本。最有解释力的比较不是“AI 对医生”,而是“契约化人机工作流对现有工作流”;必要时可设置医生单独、智能体单独、医生+智能体,以及不同交互顺序的多臂设计,以识别增益究竟来自模型、人员还是界面。

HASG 还应区分平均效应与分布效应。总体准确率提高可能掩盖少数群体受损、复杂患者被排除,或低资源机构无法复现等问题;既有案例已经表明,表面中性的算法可能因代理变量和资源分配机制而扩大群体差异^[53]。研究应预先定义性别、年龄、族群、语言、残障、社会经济状态和机构层级等亚组,并报告校准度、净获益与资源使用。对于连续照护,还需测量失访、异常升级及时性、治疗依从性和跨场景信息完整性。

4.2 共生相关伤害:识别由交互本身产生的新风险

SRH 是由人、智能体和工作流交互产生、不能仅归因于单一成员的伤害,包括自动化偏倚、错误建议采纳、责任扩散、患者自主下降、医患关系弱化、隐私侵害、工作重新分配、警报疲劳、专业能力退化、网络攻击和不公平扩大^[37,39-41,53]。错误率相同的系统可能产生截然不同的 SRH:若错误难以察觉,以高确信语言呈现,直接进入医嘱,或集中影响特定群体,其危害将远高于可被拦截的孤立错误。

SRH 评价需记录“近失事件”与恢复过程,而非只统计已造成伤害的不良事件。应建立智能体特异事件分类:目标设定错误、信息遗漏、幻觉或无依据陈述、工具调用错误、权限越界、执行失败、未及时升级、版本漂移与审计链缺失。机构应分析错误如何穿过多层防线,并以修订工作流而非简单追责个人为主要改进方式。

4.3 从报告规范到学习型证据链 研究证据应沿生命周期递进。开发与验证阶段,可用 TRIPOD+AI 规范预测模型报告,以 PROBAST+AI 评估质量、偏倚风险和适用性,并依据 CLAIM 或 MI-CLAIM 提高医学影像 AI 研究的透明度;早期现场研究应依 DECIDE-AI 重点报告人因、工作流与安全;随机试验则应依 SPIRIT-AI 与 CONSORT-AI 预先规定输入质量、版本、错误处理和人机交互;部署后还需持续

监测数据漂移、性能漂移、使用偏差与不良事件^[46-52]。对医学智能体,还应补充“行动报告最小集”:目标、记忆范围、规划策略、可用工具、权限清单、停止条件、人工干预点、失败传播路径和更新机制。

由于模型和外部工具会持续变化,传统“一篇论文对应一个冻结干预”的证据观需要扩展。建议以版本化证据档案连接研发数据、验证报告、试验结果、变更记录、部署指标与事故分析;重大变更应依预先设定的变更控制计划重新评估,小幅变更则可通过统计过程控制和哨点任务持续监测^[19]。系统越具适应性,治理越需要可重复、可审计。

4.4 成熟度与 3 种必须避免的未来 共生医学成熟

度不由智能体自主等级决定,而由 7 个问题判定:组合是否优于单独工作? 患者是否实质参与? 高风险行动是否受控? 错误是否可发现、可中断、可纠正? 责任是否可追溯? 医患关系是否增强? 公平是否改善? 一个高度自主但缺乏净获益和追责能力的系统不是成熟,而是高暴露。

必须避免 3 种未来:(1)技术替代:智能体排挤医生判断和患者参与,把复杂照护压缩为自动分流。(2)责任真空:多方共享行动,却无人对整体后果负责。(3)数字分层:高资源人群获得医生深度参与的优质共生照护,弱势人群只能接受低成本、低监督自动化。双锚点评价的伦理意义,正是使效率收益不能掩盖关系、能力与公平损失。

表 4 HASG 与 SRH 双锚点评价指标及推荐设计

领域	HASG 核心指标	SRH 核心指标	推荐比较/监测
临床结局	死亡、并发症、症状、功能、生活质量	延误、误治、可预防不良事件	务实性试验;风险调整真实世界队列
决策质量	诊断/管理适当性、校准、净获益、证据一致性	错误采纳、遗漏、过度诊疗、虚假确信	多臂人机组合试验;盲法专家评审
人因与关系	时间释放、认知负荷下降、共同决策质量、信任校准	自动化偏倚、警报疲劳、能力退化、关系弱化	交互日志、可用性研究、纵向能力评估
患者权利	知情理解、选择/退出、人工复核、申诉可达	隐性自动化、强迫接受、数据越权、无法纠错	患者报告体验;授权与申诉审计
系统韧性	升级及时、恢复时间、跨场景连续性	越权、工具故障、级联错误、审计缺口	红队测试、故障注入、哨点任务、事件复盘
公平与价值	亚组净获益、可及性、成本效果、碳与资源效率	性能差异、服务降级、数字排斥、机会成本	分层分析;预算影响;公平影响评估

HASG:人机共生增益;SRH:共生相关伤害。所有指标应预先定义临床重要差异,并报告绝对效应及不确定性。

5 未来十年的研究、治理与中国路径

5.1 从“可行”转向“值得、可控、可持续” 未来研究的首要任务,是建立可操作的定义与分类学,区分 LLM、工作流自动化、AI 智能体和智能体化 AI,明确自主性、适应性、多智能体协作与真实世界行动的最低特征^[14-15,21]。其次,应开展任务一场景特异的比较研究:不仅问系统能否给出正确答案,还要问在时间压力、信息不完整、资源约束和多病共存条件下,何种协作顺序、解释方式与接管阈值最为安全。最后,应建立 HASG 与 SRH 核心结局集,以支持跨疾病、跨机构的证据合并。

长期研究应关注两类慢变量。其一是人的变化:医生的独立推理能力、信任校准、工作意义与职业责任是否随使用而改变;患者获得的是更强的理解与控制,还是逐渐形成被动依赖。其二是系统变

化:模型更新、数据漂移、攻击方式、供应商退出和卫生政策如何改变净获益。短期模拟研究无法回答这些问题,需要开展嵌套于学习型卫生系统的多年随访。

5.2 风险分层监管与机构级临床治理 监管应围绕预期用途和实际行动风险,而不是仅围绕算法类别。用于病历摘要的同一基础模型,在生成处方、自动发送患者信息或控制设备时,风险完全不同。建议采用“任务一权限一场景”三维分级:任务决定临床后果,权限决定系统可直接造成的改变,场景决定错误被发现和补救的机会。高风险组合需实施独立验证、网络安全评估、患者参与、可解释操作记录、预先变更控制与强制部署后监测^[18-19]。

机构层面应设医学智能体治理委员会和临床智能体责任官,建立一体化的登记、准入、授权、更新、停用和事故报告制度。采购合同需明确数据权

属、日志可得性、第三方模型或工具、服务中断、版本通知、审计权与损害分担,避免关键责任被封闭供应链遮蔽。WHO、FUTURE-AI、欧盟 AI 法案和 FDA 指南提供了伦理、全生命周期与风险监管基线,SCC 则将这些基线进一步落实到床旁协作和患者权利层面^[16-20]。

5.3 教育、公平与卫生系统价值 医学教育需要把 AI 素养从选修技能升级为临床胜任力。课程应训练问题界定、证据溯源、不确定性表达、错误诱导演练、自动化偏倚、数据和算法公平、患者沟通、人工接管与事件复盘。考核既要观察使用 AI 时的协作能力,也要定期评估无 AI 时的独立能力。患者教育则应以用途和权利为中心,避免用技术术语替代真正知情。

公平必须在部署前设计。训练数据代表性只是起点,还需评估设备、网络、语言、照护者、支付和转诊资源是否使不同群体能够实际受益。机构不能以“提高效率”为由,把弱势群体默认分配到自动化通道。从四重目标到五重目标,卫生系统价值框架已将医务人员福祉和健康公平依次纳入总体目

标^[54-55];SCC 进一步要求以患者选择、人工可达和亚组净获益检验公平是否真正发生。

5.4 中国可采用“专病闭环—区域协同—家庭延伸”的渐进路径 中国拥有大型医院专科能力、广覆盖的基层网络和快速发展的互联网医疗基础,适合从高负担专病的 MSU 起步,逐步形成区域共生网络。国际随机试验已为医院级居家照护提供临床证据^[56]。在中国路径中,肺结节智能体可连接影像识别、危险分层、指南推荐、随访提醒和多学科会诊^[25,28];呼吸疾病居家居住院可连接云门诊、社区护理、家庭监测和医院升级通道^[26]。这两类路径共同说明:智能体的价值不在于替代一次诊断,而在于维持跨时间、跨场景的闭环。

实施时应坚持本地数据验证、临床知识库治理、普通话与方言可理解性、基层可用性和全国/区域监管衔接。专病智能体必须把指南证据、患者偏好和资源条件同时纳入,而非只优化疾病风险。区域扩展前应比较不同医院层级的 HASG 与 SRH,特别关注基层医生能力增强还是被动依赖、家庭照护负担是否转移、以及无智能终端人群是否被排除。

表 5 人机共生医学优先研究与治理议程(2026—2035 年)

优先领域	关键问题	推荐方法/制度	可判定里程碑
定义与分类	何谓医学智能体、智能化 AI 和有效 HITL	国际 Delphi;任务—权限—场景分级	形成可用于注册、试验和监管的统一术语
临床净获益	哪种人机组合在何种患者中优于常规照护	多臂务实性试验;MSU 分阶段研究	HASG 达到预设临床重要差异且 SRH 可接受
患者共治	持续授权、退出和人工复核如何落地	共同设计;患者报告结局/体验;权利审计	授权可理解、可撤回;拒绝者有等质替代路径
能力与人因	如何避免自动化偏倚和专业能力退化	错误诱导模拟;纵向胜任力监测	人机协作改善且无 AI 时能力不下降
生命周期安全	更新、漂移、攻击和多工具故障如何控制	版本档案;红队;哨点任务;强制事件报告	重大异常可发现、可停机、可回滚、可追溯
公平与可及	智能体是否缩小地区、语言和资源差距	亚组净获益;公平/预算影响评价	弱势群体不被默认低监督自动化服务替代
中国转化	如何贯通医院—社区—家庭和专病管理	专病 MSU;区域试点;支付与标准协同	形成可复制的安全、经济、低碳连续照护路径

HITL: 人类在回路;MSU:最小共生单元;HASG:人机共生增益;SRH:共生相关伤害。

6 结 论

医学智能体正在推动医学由人类单独认知走向共享智能,但技术能力的增长不会自动生成合宜的临床秩序。若缺少契约,智能体越能行动,患者自主、专业责任与系统安全越可能被稀释。SCC 以共享智能、有界行动、患者共治和责任追溯,把人机

共生的价值愿景转化为可检验的制度结构;FADA 闭环确保临床问题由人类界定、证据由智能体扩展、方案由多方审议、行动由有资格者授权;HASG 与 SRH 则要求任何效率与准确率收益都接受患者结局、关系质量、专业能力、公平和长期风险的共同检验。

真正先进的医学智能体,不是拥有最大自主权

的系统,而是能够在适当时行动、在不确定时求助、在越界前停止、在出错后被追溯的系统。真正成熟的人机共生医学,也不是机器替代医生或医生被动批准算法建议,而是智能共同形成、决策共同审议、行动边界清晰、患者充分参与、医生履行专业责任、机构持续监督。最终愿景是:让智能体承担信息与流程的复杂性,让医生回归判断、责任与关怀,让患者成为医疗决策的共同主体,让医疗体系由间断服务转变为连续、可学习、可信赖的健康伙伴。

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 杨达伟:素材整理、文章撰写;宋元林:素材提供、稿件审核;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] Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine[J]. *Nat Med*, 2022, 28(1): 31–38.
- [3] 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.
- [4] Kelly C J, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence[J]. *BMC Med*, 2019, 17(1): 195.
- [5] Sendak M P, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery[J]. *EMJ Innov*, 2020, 4(1):47–57.
- [6] Nagendran M, Chen Y, Lovejoy C A, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies [J]. *BMJ*, 2020, 368: m689.
- [7] Liu X X, Faes L, Kale A U, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis [J]. *Lancet Digit Health*, 2019, 1(6): e271–e297.
- [8] Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medical artificial intelligence[J]. *Nature*, 2023, 616(7956): 259–265.
- [9] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine [J]. *Nat Med*, 2023, 29(8): 1930–1940.
- [10] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. *Nature*, 2023, 620(7972): 172–180.
- [11] Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models[J]. *Nat Med*, 2025, 31(3): 943–950.
- [12] 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.
- [13] Harter S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine [J]. *EBioMedicine*, 2023, 90: 104512.
- [14] Qiu J N, Lam K, Li G H, et al. LLM-based agentic systems in medicine and healthcare [J]. *Nat Mach Intell*, 2024, 6(12): 1418–1420.
- [15] Mehndru N, Miao B Y, Almaraz E R, et al. Evaluating large language models as agents in the clinic [J]. *NPJ Digit Med*, 2024, 7(1): 84.
- [16] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance [R]. Geneva: WHO, 2021.
- [17] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models [R]. Geneva: WHO, 2024.
- [18] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [J]. *Off J Eur Union*, 2024, L 2024/1689: 1–144.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] Yang D W, Zhou J, Chen R C, et al. Expert consensus on the metaverse in medicine[J]. *Clin eHealth*, 2022, 5: 1–9.
- [23] 白春学. 元宇宙医学之我见[J]. *中国医药导刊*, 2023, 25(1):1–6.
- [24] 白春学. 医学GPT研发应用之吾见[J]. *元宇宙医学*, 2025, 2(1):6–14.
- [25] 白春学. 肺结节专家——BAIMGPT白皮书[J]. *元宇宙医学*, 2025, 2(2):55–64.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence [J]. *Nature*, 2025, 642(8067): 442–450.

- [30] McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models [J]. *Nature*, 2025, 642(8067): 451–457.
- [31] Ayers J W, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum [J]. *JAMA Intern Med*, 2023, 183(6): 589–596.
- [32] Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians [J]. *JAMA Intern Med*, 2024, 184(5): 581–583.
- [33] Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge [J]. *JAMA*, 2023, 330(1): 78–80.
- [34] 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.
- [35] 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(4): 1233–1238.
- [36] Yu F Y, Moehring A, Banerjee O, et al. Heterogeneity and predictors of the effects of AI assistance on radiologists [J]. *Nat Med*, 2024, 30(3): 837–849.
- [37] Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance [J]. *Radiology*, 2023, 307(4): e222176.
- [38] Reverberi C, Rigon T, Solari A, et al. Experimental evidence of effective human-AI collaboration in medical decision-making [J]. *Sci Rep*, 2022, 12(1): 14952.
- [39] Char D S, Shah N H, Magnus D. Implementing machine learning in health care—addressing ethical challenges [J]. *N Engl J Med*, 2018, 378(11): 981–983.
- [40] Morley J, Machado C C V, Burr C, et al. The ethics of AI in health care: a mapping review [J]. *Soc Sci Med*, 2020, 260: 113172.
- [41] Ong J C L, Chang S Y, William W, et al. Ethical and regulatory challenges of large language models in medicine [J]. *Lancet Digit Health*, 2024, 6(6): e428–e432.
- [42] Richardson J P, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare [J]. *npj Digit Med*, 2021, 4: 140.
- [43] Musbahi O, Syed L, Le Feuvre P, et al. Public patient views of artificial intelligence in healthcare: a nominal group technique study [J]. *Digit Health*, 2021, 7: 20552076211063682.
- [44] Longoni C, Bonezzi A, Morewedge C K. Resistance to medical artificial intelligence [J]. *J Consum Res*, 2019, 46(4): 629–650.
- [45] Kempt H, Nagel S K. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts [J]. *J Med Ethics*, 2022, 48(4): 222–229.
- [46] 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: 1364–1374.
- [47] 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(9): 1351–1363.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] Norgeot B, Quer G, Beaulieu-Jones B K, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist [J]. *Nat Med*, 2020, 26(9): 1320–1324.
- [53] 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.
- [54] Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider [J]. *Ann Fam Med*, 2014, 12(6): 573–576.
- [55] Nundy S, Cooper L A, Mate K S. The quintuple aim for health care improvement: a new imperative to advance health equity [J]. *JAMA*, 2022, 327(6): 521–522.
- [56] Levine D M, Ouchi K, Blanchfield B, et al. Hospital-level care at home for acutely ill adults: a pilot randomized controlled trial [J]. *J Gen Intern Med*, 2018, 33(5): 729–736.

引用本文

杨达伟, 宋元林, 白春学. 迈向人机共生医学: 医学智能体时代的共生临床契约 [J]. 元宇宙医学, 2026, 3(3): 233–242.

Yang D W, Song Y L, Bai C X. Toward human - AI symbiotic medicine: a symbiotic clinical compact for the era of medical agents [J]. *Metaverse Med*, 2026, 3(3): 233–242.