

DOI:10.61189/763560klnthj

·指南与共识·

可信医学智能体:研发、评价与临床实施中国专家共识

宋元林^{1,2,3,4},高承实⁵,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
5. 安徽栈谷科技有限公司,池州 247100

[摘要] 医学智能体正由信息生成工具演变为复杂的人工智能系统,能够感知临床状态、开展多步骤推理与任务规划、调用外部工具,并在预设权限内采取行动。因此,核心临床问题已不再是此类系统能否生成看似合理的答案,而是某一特定系统在明确权限和有效人工监督下,是否具有足够证据,可在真实临床工作流中安全行动。错误可能沿“信息—建议—医嘱—执行”路径累积并放大,因此,以准确率或考试式基准为中心的评价不足以建立临床可信性。本文综合世界卫生组织、国际医疗器械监管机构论坛等的现行指导文件,参考SPIRIT-AI、CONSORT-AI等报告与评价规范,以及近期临床研究,提出医学智能体自主性矩阵(medical agent autonomy matrix, MAAM)和医学智能体评价与证据框架(medical agent assessment and evidence framework, MAAE)。证据综合发现3类持续存在的脱节:离线能力与临床获益脱节、模型性能与系统安全脱节,以及部署前验证与全生命周期保障脱节。MAAM-MAAE通过协调风险等级、行动权限、证据成熟度和人工监督强度4个维度弥合上述差距。该框架以临床任务成功率和具有临床后果的不安全行动率为评价双锚点,从能力、安全、人机临床交互及系统全生命周期4个领域组织证据,并建立5个顺序准入门:临床必要性、系统可靠性、模拟安全性、临床可行性,以及临床有效性与可治理性。21条可操作共识条款涵盖预期用途、数据与知识治理、系统架构、工具与权限、红队测试、临床评价、机构准入、人工监督、患者保护、持续监测、版本变更、网络安全及国际标准映射。核心结论:随着自主程度、行动接近患者的程度及潜在伤害严重性增加,证据阈值、人工监督和监管审查应同步加强。现阶段,高风险医疗行动不得脱离合格卫生专业人员的实时授权和可撤销控制。

[关键词] 医学智能体;可信人工智能;大语言模型;临床评价;人机协作;风险分层;全生命周期治理;患者安全

[中图分类号] R-3 **[文献标志码]** A

Trustworthy medical agents: a Chinese expert consensus on development, evaluation, and clinical implementation

Song Yuanlin^{1,2,3,4}, Gao Chengshi⁵, 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
5. Anhui Zhangu Technology Co., Ltd., Chizhou 247100, Anhui, China

[Abstract] Medical agents are evolving from information-generation tools into complex AI systems capable of perceiving clinical states, conducting multi-step reasoning and task planning, invoking external tools, and taking actions within predefined permissions. Accordingly, the central clinical question is no longer whether such systems can generate plausible answers, but whether a given system, under explicit permissions and effective human oversight, has sufficient evidence to act safely in real-world clinical

[收稿日期] 2026-07-20

[接受日期] 2026-09-07

[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] 宋元林,博士,主任医师;E-mail: song.yuanlin@zs-hospital.sh.cn

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

workflows. Because errors may accumulate and amplify along the “information – recommendation – order – execution” pathway, evaluations centered on accuracy or examination-style benchmarks are insufficient to establish clinical trustworthiness. Drawing on current guidance from the World Health Organization and the International Medical Device Regulators Forum, reporting and evaluation standards such as SPIRIT-AI and CONSORT-AI, and recent clinical research, this paper proposes the Medical Agent Autonomy Matrix (MAAM) and the Medical Agent Assessment and Evidence Framework (MAAE). The evidence synthesis identifies three persistent disconnects: between offline capability and clinical benefit, between model performance and system safety, and between pre-deployment validation and whole-lifecycle assurance. MAAM – MAAE bridges these gaps by aligning four dimensions: risk level, action authority, evidence maturity, and intensity of human oversight. The framework anchors evaluation on clinical task success rate and the rate of unsafe actions with clinical consequences; organizes evidence across four domains—capability, safety, human – AI clinical interaction, and the system lifecycle; and establishes five sequential gates: clinical necessity, system reliability, simulated safety, clinical feasibility, and clinical effectiveness and governability. Twenty-one actionable consensus statements cover intended use, data and knowledge governance, system architecture, tools and permissions, red-teaming, clinical evaluation, institutional readiness, human oversight, patient protection, continuous monitoring, version changes, cybersecurity, and mapping to international standards. The core conclusion is that as the degree of autonomy, the proximity of actions to patients, and the severity of potential harm increase, evidence thresholds, human oversight, and regulatory scrutiny should be strengthened in parallel. At present, high-risk medical actions must not proceed without real-time authorization and revocable control by qualified health professionals.

[Key Words] medical agents; trustworthy artificial intelligence; large language models; clinical evaluation; human – machine collaboration; risk stratification; lifecycle governance; patient safety

生成式人工智能和大语言模型 (large language model, LLM) 正在推动医疗人工智能由单点预测转向目标导向的纵向任务。新兴系统可读取电子健康记录,整合影像与实验室检查结果,检索指南,生成鉴别诊断,识别潜在临床试验匹配,拟写临床医嘱,安排随访,并通过应用程序编程接口触发业务工作流。在小规模模拟肿瘤病例系列中,多模态智能体在工具选择和综合决策方面显示优势;随机研究亦提示,嵌入经专门设计的协作界面的 LLM 可改善临床医师对复杂病例的推理。然而,这些发现不能外推为常规临床中自主诊断或治疗的证据^[1-9]。医学智能体带来的关键变化,并非其输出更像医生,而是输出可能被转化为行动。因此,多步骤轨迹中的局部错误、恢复失败或权限误用,均可能传播为系统层面的伤害。

现有证据存在3处断裂。(1)离线能力与临床获益脱节:考试题、单轮回答和回顾性数据集很少再现信息不完整、时间压力、工作流中断及人机交互,可能造成“评价充分”的错觉^[3-6,10]。(2)模型性能与系统安全脱节:风险由基础模型、检索知识、记忆、规划器、工具、界面、权限和使用者共同产生;仅验证模型,并不等于验证“临床医师—智能体—工作流”系统。(3)部署验证与全生命周期安全脱节:基础模型、知识库、提示模板和界面持续变化,数据分布、临床路径及使用者依赖程度也会漂移;一次性验证无法保证持续安全^[11-26]。因此,适当的评价单元不是孤立模型,而是由特定版本、权限集和临床

工作流共同定义的完整系统。

WHO将透明、问责、包容、公平和持续治理视为卫生健康人工智能的基础要求。FDA关于人工智能医疗器械变更控制的框架、IMDRF关于医疗软件临床评价的框架、欧盟对高风险人工智能的要求,以及ISO/IEC风险与管理体系,均趋向于将预期用途、风险管理、证据、监测和变更控制连接为全生命周期路径^[11-26]。中国已形成覆盖医疗器械软件、人工智能医疗器械、网络安全、个人信息保护和生成式人工智能服务的监管基础,并发布涵盖84个卫生健康行业应用场景的参考指引^[27-32]。然而,能够规划并调用工具的医学智能体超出了传统单模型评价范畴。目前仍缺少一个可将国际标准与中国临床实践、分级诊疗、互联网医院和居家医疗相结合的可操作框架。

为弥合上述差距,本共识提出成对的MAAM-MAAE框架。MAAM将风险等级(风险I~IV)与行动权限(A1~A5)结合,用于界定系统可以做什么以及在何种条件下可以做;MAAE采用从技术验证、临床结局到持续治理的渐进式证据,用于判定何时已有充分证据支持相应行动。人工监督和5个准入门将两套框架连接为闭环:任务明确、安全内生、证据渐进、受控部署和持续治理。本框架亦从元宇宙医学、医学GPT、BAIMGPT、云门诊及居家医院路径中提炼情境经验,但不将其视为临床有效性的证明。与边界不清的通用自主医疗相比,本框架更支持依托本地知识基础设施和医患协作的专病、多模态

系统^[33-39]。

1 范围、定义与核心原则

1.1 医学智能体的定义与边界 医学智能体是面向明确医学目标的人工智能系统,能够感知患者及照护环境,将医学知识与个体状态相结合,开展多步骤推理和任务规划,调用一个或多个外部工具,在预设权限内生成建议或执行行动,并根据行动结果调整后续行为。其最小功能链包括感知、推理、规划、行动和反馈;每一组成部分均应可约束、可记录、可验证。仅执行单一分类、预测、摘要或问答任务,且不维持任务状态、不调用工具或不根据反馈进行调整的系统,应归类为人工智能模型或辅助工具,而非医学智能体。多智能体架构并不会天然提高可靠性;通信失真、共识偏倚、责任归属和级联失败均需单独验证。

自主并不意味着无需问责。医学智能体的自主性应拆分为6项可审计权限:任务选择、计划生成、工具选择、参数生成、触发执行和结局复核。临床风险取决于错误能否越过人工审核并到达患者,而不取决于产品标签或参数规模。同一基础模型用于文献检索与用于胰岛素剂量调整时,应适用显著不同的监管分类、证据阈值和监督强度。

共识条款1(强推荐;证据确定性:中等) 医学智能体必须依据预期用途和完整行动链进行定义,而不能仅依据基础模型、参数规模或商业名称中的“智能体”表述。研发和准入文件至少应明确目标患者、应用场景、预期使用者、输入、输出、工具、行动、反馈、禁止用途及承担责任的临床主体。

1.2 可信医学智能体的可验证属性 可信性不是抽象声誉,而是必须由证据证明的系统属性。可信医学智能体应具备7项属性:临床有效性;安全性与韧性;透明性与可追溯性;公平性与可及性;隐私与安全;可用性与可监督性;以及问责性与可治理性^[11-13,18-26,40-41]。这些属性之间可能相互竞争;增加解释可能提高认知负担;扩大数据覆盖可能增加隐私风险;开放更多工具访问权限可能提高任务完成率,同时扩大攻击面。因此,评价应说明每项权衡的临床依据并刻画残余风险,而不能用综合评分掩盖关键薄弱环节。

可解释性应由“生成理由”转向“提供可核验的证据链”。系统应区分患者事实、模型推断、外部证据和建议行动,并记录数据时间戳、知识版本、工具输出、规则校验结果及人工修改。LLM生成的自然语言推理并不能揭示模型内部的因果过程,也不能

替代安全性证据。对于高风险任务,来源可追溯性、规则一致性及对关键变量的敏感性,通常比冗长的思维链更具临床信息价值^[3,6,42-43]。

共识条款2(强推荐;证据确定性:中等) 可信性属性必须转化为预先设定的指标、阈值、测试场景和问责记录。未经验证的“可解释”“公平”或“安全”声明不得作为证据。

1.3 五项核心原则 本共识提出5项核心原则。(1)以临床需求驱动研发:系统应解决现有路径中对患者重要且可测量的缺口,而不应仅因技术可行便创造低价值需求。(2)证据与治理应风险适配:自主程度、行动接近患者的程度、伤害严重性、可逆性及可检测性共同决定所需证据和监督强度。(3)将安全内置于设计:最小权限、适当拒答、基于规则的护栏、自动停止、人工接管和可审计日志应在架构阶段纳入。(4)人和机构继续承担责任:卫生专业人员和医疗机构保留关键决策权、最终否决权及连续照护责任。(5)治理贯穿全生命周期:部署并非研发终点;持续监测、事件学习、版本控制和定期再认证均属于产品治理。

共识条款3(强推荐;证据确定性:中等) 所有医学智能体均应遵循临床需求导向、风险适配、安全内生、人工问责和全生命周期治理原则。当各原则发生冲突时,应优先保护患者生命、基本权利和公平可及。

2 共识制定与推荐形成方法

本共识采用结构化范围综述与标准映射方法。证据检索于2026年7月15日完成,优先纳入国际组织和监管机构文件、方法学指南、高质量综述、临床研究、智能体安全研究及现行中国标准,共保留66篇核心参考文献。报告参考RIGHT,证据确定性和推荐强度参考GRADE。鉴于软件快速演进,另行考虑工程一致性、可重复性、临床直接性及版本时效性^[40,42,44-53]。高确定性是指获得多中心前瞻性研究的直接支持或一致的强制性标准支持;中等确定性是指一致的前瞻性、真实世界或高质量方法学证据;低确定性主要来源于回顾性、模拟或小样本研究;极低确定性主要来源于机制推断和专家意见。强制性标准可确立治理义务,但不能构成临床有效性证据。

推荐分为强推荐和条件性推荐。强推荐表示在多数目标场景中,获益—风险平衡、可行性和伦理依据足够明确,可作为最低要求;条件性推荐则要求结合风险等级、资源、价值与偏好以及本地验

证作出决定。医学智能体以患者重要结局为终点的证据仍然有限。因此,本共识可依据预防原则,对可逆、低负担的最低安全要求提出强推荐,而效果阈值和更新间隔仍应依据具体场景确定。符合标准并不等同于具有临床有效性;后者需要前瞻性比较评价。

本共识包含21条共识条款,尚需经过2~3轮德尔菲调查、正式共识会议、公开征求意见和国际外部评审。参与者应包括医学、护理、药学、患者代表、人工智能、方法学、统计学、人因工程、伦理与法律、网络安全、监管和卫生经济学专家。最终投稿应报告专家组构成、利益冲突管理、各轮应答率、评分分布、共识阈值和修订轨迹;不得事后重构投票率。建议预先规定:同意或强烈同意比例 $\geq 80\%$,且5分制评分中位数 ≥ 4 分,判定为达成共识;70%~79%为需要修订后重新审议; $< 70\%$ 则删除。在上述程序完成前,本文应视为供评审的循证框架,而非已完成正式投票的最终专家共识。

共识条款4(强推荐;证据确定性:中等) 医学智能体相关共识、指南及评价研究应采用透明、可重复的方法,披露证据检索、利益冲突、患者参与及推荐形成程序,并根据研究阶段采用SPIRIT-AI、CONSORT-AI、DECIDE-AI、TRIPOD+AI、PROBAST+AI、STARD-AI、CLAIM、MI-CLAIM-GEN和CHEERS-AI等适用规范。

3 可信医学智能体的研发与评价标准

3.1 从临床任务到受控架构 研发应始于临床任务声明,而非模型选择。任务声明应明确目标人群、照护场景、现有路径、未满足需求、预期获益、最严重可信伤害、替代方案、禁止范围及成功标准。任务边界还应规定时间、地点、人员资质、数据完整性和资源条件。例如,肺结节随访系统应区分筛查人群与既往肿瘤、免疫抑制、妊娠、急性症状或儿童患者;同时应明确何时适用本地指南,何时需要转诊或拒答^[36,39]。预期用途发生实质性变更,等同于形成新的产品假设,必须重新开展风险分析。

数据治理应覆盖合法来源、代表性、时效性、标注可靠性、缺失机制、数据泄漏及亚组覆盖。训练、调优、测试和外部验证数据应按患者、机构和时间进行适当隔离。重复记录、同一患者的多模态观察以及模板化文本均可能造成隐性泄漏。中文医学缩写、否定表达、复制文本、时间错配及跨机构编码差异需要专项管理。参考标准应基于与任务相适应的专家判断或临床结局,并报告裁决程序;历史

医嘱不应未经审慎分析便作为金标准^[5,43,47-50]。

知识治理应对指南、药品目录、处方信息、共识、机构政策和患者特异性信息分层管理。检索增强生成(retrieval-augmented generation, RAG)可缓解静态知识过时,但不能保证正确性。检索召回不足、来源冲突、文本切片失真、引文错配及知识库污染均会引入新的失效模式。系统应维护来源白名单、适用性标签、生效与失效日期、冲突解决优先级及撤回机制,并验证完整的“检索—解释—应用—引证”链^[3-6,38,54-55]。

架构文件应明确基础模型及版本、系统提示词、记忆类型及保留期限、规划器、工具清单、接口、身份与权限、规则引擎、审计日志和回退路径。高风险参数应由确定性代码或经过验证的计算器处理,而不应依赖自由文本生成。工具调用需要进行模式校验、输入和输出类型检查、幂等性控制及事务确认。对医嘱录入、处方、设备或消息系统的写权限,应与读权限在物理或逻辑上分离。

共识条款5(强推荐;证据确定性:中等) 研发前,团队必须形成具有版本控制的临床任务声明和危害分析,明确预期用途、禁止范围、替代路径、最严重可信伤害、CTS和CCUAR目标以及停止规则。

共识条款6(强推荐;证据确定性:中等) 配置清单和来源链必须覆盖数据、知识、模型、提示词、记忆、工具和工作流。任何不可追溯的关键组件均不得用于风险Ⅲ~Ⅳ级临床任务。

共识条款7(强推荐;证据确定性:中等) 医学智能体应实施最小权限、工具白名单、参数校验、适当拒答、异常自动停止、即时人工接管、端到端日志及故障安全回退。A4~A5行动的授权控制必须独立于语言模型。

网络安全是临床安全的组成部分。提示注入可通过外部文档向RAG引入恶意指令;数据投毒可在总体基准性能无明显下降的情况下,选择性改变医学建议;记忆污染或工具劫持则可能跨会话持续存在。现有通用防御仍不完善,仅依赖提示词拒绝并非可靠控制措施^[19-20,54-56]。测试应覆盖直接和间接提示注入、未授权工具调用、敏感数据泄露、供应链依赖、拒绝服务、模型提取、数据投毒、对抗性输入、身份冒充和审计绕过。应由独立团队开展红队测试。

共识条款8(强推荐;证据确定性:中等) 部署前,系统必须接受与风险相称的安全威胁建模、渗透测试和红队测试。当可重复攻击路径能够导致未授权行动、高风险错误或数据泄露时,应暂停准

入,直至控制措施有效且完成回归测试。

3.2 四领域评价体系 第1领域为能力。评价应测试状态识别、信息整合、推理、规划、工具选择、参数生成、行动执行、反馈利用和错误恢复。评价单元应由单一答案扩展至完整任务轨迹,并报告分步骤和端到端成功率、重试次数、延迟及成本。测试集必须包含信息不完整、相互矛盾、超出适用范围及无唯一答案的病例,以评价澄清、拒答和升级能力。仅使用LLM作为裁判可能产生同源偏倚;关键临床结局应由盲法专家或可验证规则进行裁决。

第2领域为安全。核心关注不应是错误总数,而应是错误的临床后果。危害分类应包括事实错误、漏识禁忌证、升级延迟、不必要检查、剂量错误、未授权行动、隐私泄露、歧视性输出及系统性级联失败。每一事件均应描述严重程度、可逆性、暴露持续时间、是否到达患者、是否被人工拦截以及是否属于未遂事件。较好的平均安全评分不能抵消罕见灾难性事件;预先规定的不可接受伤害应直接导致相应准入门不通过。

第3领域为人机临床表现。研究单元是“临床医师—智能体—工作流”系统,比较对象应为当前最佳工作流,而不是理想化的无辅助环境。评价应考虑临床医师独立判断与AI建议的呈现顺序、分歧的可见性、人工修改、自动化偏倚、认知负担、任务时间、患者体验及临床结局。随机研究显示,辅助方式和界面设计可改变临床医师决策及表现^[7-8,57]。另有证据^[58]表明,AI错误可能使部分放射科医师表现恶化,传统经验指标也不能可靠识别谁会获益。这些发现支持开展本地模拟和个体化培训,而不能假定AI辅助对所有人员均有益。

第4领域为系统与全生命周期表现,包括可用性、互操作性、停机韧性、成本效益、环境影响、可及性、公平性、性能漂移、知识过时及长期安全。实施研究除采纳率外,还应报告适宜性、可接受性、实施忠实度、可持续性和非预期后果。经济学评价应纳入系统集成、人工审核、监测、更新、事件管理和退役成本,而不能只计算模型调用费用^[40,51,59-62]。

共识条款9(强推荐;证据确定性:中等) 医学智能体必须从能力、安全、人机临床表现及系统全生命周期表现四个领域进行评价。低层级领域的证据不能替代高层级领域证据;实验室基准成功不能证明临床获益。

共识条款10(强推荐;证据确定性:中等) 所有关键评价均应按预期人群、机构、设备、语言、年

龄、性别及临床相关脆弱亚组报告性能和安全性。总体均值不得掩盖不可接受的亚组风险。

3.3 双锚点结局与渐进式证据 CTS是指在预先规定的时间、资源、权限和质量约束内完成临床任务的比例。成功必须包括关键步骤完成、最终结果可接受及必要升级得到执行。若行动虽已完成,但漏识禁忌证或未传达关键结果,不得判定为成功。CCUAR是指在预先规定且可评价的行动机会中,智能体提出或执行的行动若未被拦截,可能造成具有临床意义伤害的比例。分析单位——病例、决策机会或行动——应预先设定。分子应由盲法临床专家依据严重度量表裁决,并区分被人工拦截的未遂事件与实际到达患者的事件。

双锚点可防止评价只重视获益而仅简单计数错误,或相反只重安全而忽视临床价值。高CTS不能补偿不可接受的CCUAR,极低CCUAR也不能证明临床价值。报告还应包括净临床获益、严重错误率、适当拒答率、人工接管率、适当升级时间、校准度、最差亚组表现和不确定性区间。对于诊断或预测任务,敏感度、特异度和受试者工作特征曲线下面积不能替代校准或决策曲线分析;对于生成任务,BLEU或ROUGE等文本相似性指标不能替代事实准确性或临床后果评价^[5,42-43,47-50,63-64]。

证据应依次推进:内部技术验证、独立外部验证、真实数据回顾性验证、前瞻性静默部署、有监督早期临床评价、比较效果研究及部署后真实世界监测。静默部署可在不影响患者的情况下发现数据接口、时间戳、缺失和工作流不匹配。早期临床评价主要测试可用性、安全护栏和人机交互。风险Ⅲ~Ⅳ级系统在常规使用前,应在比较研究中证明可接受的净获益,而不能仅依赖模型准确性不劣效^[7-8,42,44-51,58-60]。该序列是渐进式而非机械线性;版本、场景或权限发生实质性变化时,系统应回到相应验证阶段。

共识条款11(强推荐;证据确定性:中等) 所有医学智能体研究均应预先规定CTS和CCUAR为核心结局,并报告95%置信区间、严重度分层、人工拦截及患者暴露情况。仅报告准确率、任务完成率或满意度不足以支持评价。

共识条款12(强推荐;证据确定性:中等) 风险Ⅱ级及以上系统至少应完成独立外部验证和前瞻性静默部署。风险Ⅲ级系统应接受有监督的前瞻性临床评价。原则上,风险Ⅳ级系统应具有比较效果证据、长期安全监测及独立数据和安全监督。

表 1 四领域评价体系及推荐核心结局

评价领域	评价单元	核心结局	优选研究设计	不可替代的判断
能力	完整任务轨迹	CTS; 步骤成功率; 适当拒答; 错误恢复; 延迟与成本	独立测试; 外部验证; 边界病例	端到端任务成功; 不能以文本相似性替代
安全	危害与控制链	CCUAR; 严重错误; 未授权行动; 未遂事件; 人工拦截; 隐私/安全事件	沙箱; 故障注入; 独立红队测试	任何预设不可接受伤害均可否决准入
人机临床表现	临床医师—智能体— workflow 系统	决策质量; 患者结局; 认知负担; 自动化偏倚; 时间; 体验	静默部署; 有监督前瞻性研究; 随机或准实验设计	与当前最佳工作流比较, 并评价呈现顺序
系统—全生命周期	机构与纵向运行	采纳; 实施忠实度; 可持续性; 漂移; 公平性; 成本效益; 环境影响	真实世界监测; 实施与卫生经济学评价	纳入集成、审核、监测、更新、事件和退出

每一领域均应报告系统版本、目标人群、95%CI 及临床相关亚组。不能根据低层级领域结果推断高层级领域证据。

4 风险分层与临床准入

4.1 四级风险分层 风险 I 级为信息型系统: 主要读取、整理或解释信息, 错误通常在使用前接受审核, 例如文献检索、病历摘要和教育。风险 II 级为建议型系统: 输出可能影响临床判断, 但不直接改变医嘱, 例如风险预警、鉴别诊断和检查建议。风险 III 级为工作流型智能体: 可拟写医嘱、处方、检查、转诊或随访计划, 或自动执行能够实质改变照护流程、但可逆的数字任务。风险 IV 级为行动型智能体: 可影响药物剂量、生命支持参数、介入操作、设备控制或其他可能造成严重且不能迅速逆转伤害的行动。

风险等级不应仅依据功能名称确定, 而应综合伤害严重程度与发生概率、可检测性、可逆性、患者脆弱性、时间敏感性及与人工审核的距离。面向公众的分诊建议可能仅为文本, 但若可能延误急诊救治, 至少应按风险 II 级管理。自动取消检查可能因其工作流后果而达到风险 III 级。多功能系统应按当前启用功能中的最高风险等级准入; 若停用高风险模块, 机构可根据实际部署边界重新评估风险。

共识条款 13(强推荐; 证据确定性: 中等) 医学智能体应按风险 I 级信息型、风险 II 级建议型、风险 III 级工作流型或风险 IV 级行动型系统管理。风险等级应在预期用途文件、用户界面和日志中明确标示, 并在功能、场景或权限变更后重新评估。

4.2 五级行动权限与人工监督 A1 允许读取、整理和解释信息; A2 允许预警、建议和风险沟通; A3 允许拟写医嘱、处方、检查或随访计划, 但不得提交; A4 允许在白名单和可撤销范围内执行数字任务; A5 允许改变治疗、设备或人体状态的高风险行动。

权限应授予特定工具、参数和环境; 笼统授权“访问电子健康记录”不可接受。身份认证、授权和审计必须由确定性安全层执行, 模型只能在该边界内提出行动请求。

人工监督必须有效, 而不能停留于名义。有效监督要求监督者具有识别错误并阻止行动所需的时间、信息、能力和权限。频繁预警、默认接受和无差别确认可能增加警报疲劳和自动化偏倚。关键风险 II 级建议应具有可核验来源, 并由使用者主动采纳。任何改变工作流的风险 III 级行动均须由合格专业人员确认。风险 IV 级要求实时监督、关键参数双重确认、异常自动停止和即时接管。临床医师保留否决权, 但机构不得将系统设计缺陷的责任转嫁给个体使用者。

共识条款 14(强推荐; 证据确定性: 中等) 行动权限应依据最小权限原则, 由 A1 向 A5 逐级授予。A5 行动必须由合格卫生专业人员在充分知情且可撤销的基础上实时授权或监督。现阶段不推荐无监督的自主 A5 行动。

4.3 五个临床准入门 五个准入门将渐进式证据链转化为可操作流程。准入门 1“临床必要性”确认问题具有重要性、目标可测量, 且系统能够改善或补充现有路径而不促进低价值医疗。准入门 2“系统可靠性”评价数据、知识、模型、工具、权限、网络安全和质量体系。

准入门 3“模拟安全性”要求通过标准病例、沙箱、边界场景、故障注入及红队演练。准入门 4“临床可行性”要求开展外部验证、静默部署及与风险相称的早期临床研究。准入门 5“临床有效性与可治理性”要求具有可接受的净获益, 并建立监测、问责、变更控制和退出安排。



图1 MAAM-MAAE:风险、权限、证据与人工监督的对应关系

准入门用于约束证据要求,而不是行政检查表。较早准入门未通过时,不能因商业压力、专家印象或个别成功而绕过。条件性准入应明确持续时间、使用量、患者范围、监督方式和停止规则。部署地点、基础模型或关键工具发生重大变化时,系统应返回相关准入门,而不是仅进行界面测试。监

管授权可能是部署的必要条件,但不能替代医疗机构对本地数据、工作流、人员和风险控制的验证。

共识条款 15(强推荐;证据确定性:中等) 医学智能体应依次通过5个临床准入门,并保留可审计的决策记录。实验室基准结果不能绕过准入门3~5而支持常规临床使用。

表2 风险等级、行动权限、最低证据与人工监督的对应关系

风险等级	典型任务	允许行动	最低准入证据	最低人工监督
风险 I 级/信息型	检索;摘要;教育	A1;有限 A2	技术验证+独立外部验证;验证事实性、隐私与可用性	抽样审核;明确排除诊断/治疗;快速纠正
风险 II 级/建议型	风险预警;鉴别诊断;检查建议	A1~A2	真实数据外部验证+前瞻性静默部署;验证校准、拒答和亚组	关键建议须主动采纳;来源可核验;临床医师独立判断
风险 III 级/工作流型	拟写医嘱/处方;检查预约;随访调整	A1~A3;白名单 A4	有监督前瞻性评价;与最佳工作流比较安全、效率和实施结局	合格专业人员确认每项改变工作流的行动;可立即撤销并回退
风险 IV 级/行动型	剂量调整;生命支持;介入或设备控制	A1~A4;受控 A5	比较效果研究+长期监测+独立安全监督;严格监管和本地再验证	实时监督;关键参数双重确认;自动停止;即时接管;禁止无监督 A5

A1为信息处理,A2为建议,A3为拟写,A4为受控执行,A5为高风险行动。多功能系统按当前启用功能中的最高风险等级准入。

5 临床实施与全生命周期治理

5.1 机构准入与实施 医疗机构应建立医学智能体治理委员会,至少包括医学、护理、药学、信息技术、临床工程、质量与安全、伦理、法律、网络安全及患者代表。委员会负责产品清单、风险分类、本地验证、权限审批、培训、事件管理和退役。采购合同应确保机构能够获得版本信息、关键性能数据、事

件通知、数据使用条款、分包商信息、服务终止方案和审计材料。

本地验证应使用连续或具有代表性的机构数据,并覆盖常见病例、具有临床意义的罕见病例、脆弱亚组、夜间和节假日工作流、接口故障及数据缺失。部署应按沙箱、静默运行、限定使用者、限定患者和扩大使用的受控阶段推进,同时保留平行人工流程和故障安全回退。

共识条款 16(强推荐;证据确定性:中等) 部署前,医疗机构必须完成本地性能与安全验证、伦理与监管审查、接口和网络安全测试、责任分配、人员培训、患者告知、停机预案及持续监测计划。未达到预设阈值的系统,不得以“试用”为名影响患者照护。

表 3 医学智能体部署的机构最低准入核查表

序号	治理领域	最低问题	应保留证据
1	临床价值	是否存在重要未满足需求、明确目标及当前最佳比较方案?	本地临床任务声明;基线工作流;目标阈值
2	监管与伦理	是否已明确医疗器械属性、研究属性、数据义务及生成式 AI 义务?	监管判定;伦理/合规意见;利益冲突
3	本地性能	是否评价连续数据、关键亚组、夜间/节假日工作流及数据缺失?	锁定方案;结果;失败病例;签署结论
4	集成与安全	是否验证身份、权限、接口、日志、停机和攻击面?	接口验收;威胁模型;红队与回退记录
5	人工监督	监督者能否理解、质疑、停止和逆转系统,并在停机时继续照护?	责任矩阵;培训考核;停机演练
6	患者权利	告知、拒绝、替代方案、隐私及伤害报告是否可及?	分层告知;同意/异议和投诉路径
7	分阶段部署	部署是否按沙箱、静默、限定范围和扩大使用推进?	阶段计划;通过/修订/停止记录
8	监测与事件	能否及时追踪 CTS、CCUAR、亚组、漂移和事件阈值?	仪表盘;分级响应;通知与根因分析模板
9	版本与退出	供应商更新、停止服务或合同终止后,机构能否锁定、回滚和迁移? 变更计划;版本清单;退出和连续性计划	

任何涉及严重患者安全、重大隐私泄露、监督无效或缺乏回退路径的关键项目未通过,均应作出“停止”或“修订”决定。

5.2 人工监督、能力维持与问责 在诊断和治疗中,优选的协作顺序为:独立判断、结构化比较、明确呈现分歧和人工裁决。在独立评估前先接触 AI 结论可能使临床医师产生锚定,而先行独立判断可能有助于保留批判性推理;最佳顺序仍取决于具体任务并需验证。2026 年一项随机试验^[8]发现,无论 AI 作为第一意见还是第二意见呈现,面向协作设计的界面均提高了病例层面的诊断准确率,但并未证明真实世界患者结局改善。因此,界面和工作流设计应视为干预组成部分。

长期依赖可能导致技能衰退、警觉性下降和责任扩散。机构应通过无 AI 演练、抽样盲法测试、关键技能再认证和停机演练维持专业能力。

共识条款 17(强推荐;证据确定性:中等) 人工监督的有效性应通过错误拦截、分歧管理、覆盖质量、响应时间和停机应对能力加以证明。被动确认不构成充分监督,“临床医师最终负责”的表述不能替代系统和机构问责。

5.3 患者保护、信息告知与公平性 患者告知深度应与风险和影响相称。低风险、后台文书支持可通过机构层面的透明度声明进行告知。对于直接与患者交互,或影响分诊、诊断、治疗和资源配置的系统,应告知患者是否使用 AI、AI 执行何种任务、临床医师如何监督、数据如何使用、患者能否拒绝、有哪些替代方案以及如何报告错误。风险 III ~ IV 级或研究用途应进行明确的个体告知,或采用经伦理批准的知情同意安排。在适用情况下,AI 生成内容还应按照中国现行规定进行标识,并使用符合临床情

境、患者可理解的语言^[30-32]。

公平性评价不能止于是否使用敏感变量。训练数据代表性、医疗服务可及性、语言和数字素养、历史治疗偏倚、标签质量及部署资源均可能影响结局。综述^[65-66]显示,临床 AI 公平性研究仍集中于有限属性,且很少在临床工作流中评价干预。因此,评价者应报告最差亚组表现、错误严重度差异及服务拒绝率。若某一亚组代表不足,应限制预期用途并明确不确定性,而不能以总体表现加以掩盖。

共识条款 18(强推荐;证据确定性:中等) 对于直接与患者交互或实质影响照护的智能体,机构应提供风险分层的透明告知、退出及替代路径、隐私保障和伤害报告渠道。部署不得削弱患者获得合格卫生专业人员照护的权利。

5.4 持续监测与安全事件学习 部署后监测仪表盘至少应追踪 CTS、CCUAR、严重错误、适当拒答、人工接管、未完成任务、延迟、停机、知识过时、版本分布、亚组差异、网络安全事件、使用者投诉及临床结局。监测阈值应结合绝对限值、相较基线的变化和统计过程控制;低使用量场景可采用累积事件和专家评审。监测数据应关联具体版本、机构、使用者类型和场景,以区分模型漂移、工作流变化和培训不足。

安全事件包括造成伤害的不良事件、未遂事件、潜在系统缺陷和网络安全暴露。发生严重事件后,应立即限制或暂停相关功能,保存日志及输入输出,并启动临床处置和根因分析。完成整改后,系统必须针对索引事件、相关场景及回归测试集重

新测试。

共识条款 19(强推荐;证据确定性:中等) 部署后监测必须主动开展、版本可追溯,并与分级事件响应相连接。死亡、永久性伤害、潜在人群级风险、未授权高风险行动或重大数据泄露,应触发立即停用、调查、通知、整改和重新验证。

5.5 版本、更新与退出治理 医学智能体的版本由基础模型、权重或服务版本、系统提示词、RAG 语料库快照、嵌入模型、规划器、记忆策略、工具接口、规则集和工作流配置共同定义。任一组成部分发生变化均可能改变性能;基础模型不变并不能证明产品未变。变更可分为轻微、重要或重大。仅涉及格式或不影响临床逻辑的修复可进行回归测试;知识库、提示词或工具参数变更需验证受影响任务;基础模型、预期用途、行动权限或核心工作流变更通常属于重大更新。预定变更控制计划应明确允许变更的范围、数据与验证方法、影响评估、审批责任和部署后监测,并与 FDA 2025 年最终 PCCP 指南及 ISO/IEC 管理体系相协调^[15,21-26]。

共识条款 20(强推荐;证据确定性:中等) 模型、知识、提示词、记忆、工具和工作流的任何变更均应纳入变更控制。重大更新应重新开展风险分析、性能评价和临床影响评价,并返回与其影响相适应的准入门。

6 中国要求与国际标准的衔接

中国治理不应简单复制国际文件,而应在共同国际底线基础上增加符合中国场景的要求。共同底线包括质量管理、医疗器械风险管理、软件生命周期过程、人工智能风险与管理体系、临床评价、透明报告、网络安全及高风险人工智能基本权利保护^[11-26,40,42,44-53]。中国场景的增量要求包括超大规模且高度分级的医疗体系、中文自由文本和本地术语、互联网医院、区域平台、居家医疗、国产基础模型、跨机构数据合规、基层资源约束和多中心真实世界评价^[27-39]。

医学智能体在中国是否属于医疗器械及其分类,应依据预期用途、功能和现行法规确定。用于诊断、治疗决策或设备控制的产品,研发者应主动与药品监督管理部门沟通。即使系统不纳入医疗器械路径,医疗机构仍应依据风险实施相称的质量与安全治理。生成式人工智能、数据安全、个人信息保护、算法备案和 AI 生成内容标识等横向要求,应与医疗垂直监管协同;符合某一领域要求并不能免除其他领域义务^[27-32]。

建议建立国家或区域医学智能体注册与评价网络。共同最小数据集应记录版本、预期用途、风险 I~IV/A1~A5 分类、CTS、CCUAR、亚组表现、事件和变更。受控沙箱、标准化病例库和独立红队平台应支持在三级医院、基层医疗机构和居家场景中开展分层验证。BAIMGPT 专病闭环和云门诊支持的“医院—社区—家庭”路径经验提示,价值评价应覆盖连续性、可及性、环境足迹、社会效益和经济价值,而不能只关注单一医院效率^[36-39]。

共识条款 21(强推荐;证据确定性:中等) 中国应建立与国际标准衔接并纳入本地临床路径和真实世界数据的医学智能体治理体系。该体系应提供研发、评价、机构准入、事件报告、版本变更及国际映射的共同工具,并优先支持高价值专病和基层可及性应用。

综上所述,可信医学智能体的临床价值不能由模型规模、单次演示或问答准确率推断。可信性的最小评价单元是“临床医师—智能体—工作流”系统;证据的最小评价单元,是特定版本在特定患者人群、任务、权限集和监督水平下的获益—风险特征。MAAM 通过协调风险与行动权限约束系统“可以做什么”,MAAE 则通过协调权限与证据成熟度界定系统“何时可以行动”。CTS 和 CCUAR 防止能力与安全相互替代,5 个准入门则把证据转化为可暂停、可修订、可审计的临床决策。

与完全自主照护相比,当前证据更支持专病、多模态、证据约束、权限受控且持续治理的人机协作系统。研究应由模拟病例走向多中心临床工作流,由单次输出走向纵向任务轨迹,由平均准确率走向患者重要结局和最差亚组表现,由初步可行性走向长期安全、经济价值及环境影响。现有证据仍主要由标准、方法学研究、模拟研究和少量前瞻性研究构成,异质性较大,真实世界患者结局、长期伤害及成本效益证据不足。因此,本框架界定的是风险适配的最低治理要求,不应被解释为任何特定产品已经具有临床有效性或自主行动能力。

本共识下一阶段应开展多利益相关方德尔菲过程和正式共识会议,随后公布投票结果、证据表、研发与机构准入核查表、安全事件报告表及版本变更控制表等在线附录。本共识建议每 2 年更新 1 次;如发生重大监管变更、基础模型能力显著进展或严重安全事件,应提前更新。只有在临床任务明确、架构透明、风险与权限匹配、证据充分、人工监督有效、患者权利得到保护且治理贯穿全生命周期时,医学智能体才适合进入真实世界医疗。

表 4 中国医学智能体治理框架与国际标准的功能映射

治理功能	主要国际依据	本共识的操作要求	中国场景增量要求
伦理与基本权利	WHO 2021/2024; 欧盟 AI 法案	透明、问责、公平、患者参与及分层告知	可理解的中文告知; 互联网医院及数字弱势群体的替代方案
医疗软件临床评价	IMDRF N41; SPIRIT/CONSORT/DECIDE-AI	渐进式分析有效性、科学有效性和临床表现; 五个准入门	跨机构层级静默部署, 并依据本地路径验证
质量与风险管理	ISO 13485; ISO 14971; ISO/IEC 23894	临床任务声明、危害分析、残余风险和全生命周期风险文件	与 NMPA 人工智能医疗器械、软件及网络安全审评衔接
软件与 AI 管理体系	IEC 62304; ISO/IEC 42001	配置清单、版本、验证、发布、维护、退役和管理责任	国产基础模型、区域平台及多供应商责任矩阵
更新与变更	FDA PCCP 2025	预定范围、验证、影响评估、审批、回滚和监测	覆盖模型、知识、提示词、工具及机构路径变更
报告与偏倚	TRIPOD+AI; PROBAST+AI; STARD-AI; CLAIM	外部验证、校准、适用性、偏倚风险及最差亚组表现	中文语料、基层数据、区域疾病谱及可及性
生成式 AI 评价	MI-CLAIM-GEN; FUTURE-AI	模型卡、任务边界、事实性、可追溯性、可用性、鲁棒性及环境影响	专病知识基础设施、医学术语及版本化本地指南
网络安全	NIST AI RMF/GenAI Profile; IEC 81001-5-1	对提示注入、投毒、记忆、工具、供应链和敏感数据开展红队测试	与等级保护、跨境数据、个人信息及机构身份控制协同
生成内容治理	WHO LMM 指南	披露 AI 身份、内容来源、异常报告及患者申诉	与生成式 AI 暂行办法及 AI 生成合成内容标识规则衔接
实施与价值	CHEERS-AI; 实施科学框架	与最佳工作流比较; 纳入总成本、可持续性、公平性及环境足迹	医院—社区—家庭照护、云门诊服务及基层可及性

功能映射不代表法律等效或自动合规。产品属性、分类和审批路径仍应遵循中国现行法律法规。

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 宋元林: 素材提供、稿件审核; 高承实: 素材提供、文章修改; 杨达伟: 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] Ferber D, El Nahhas O S M, Wölflein G, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology[J]. Nat Cancer, 2025, 6 (8): 1337-1349.

[2] 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.

[3] Teo Z L, Thirunavukarasu A J, Elangovan K, et al. Generative artificial intelligence in medicine[J]. Nat Med, 2025, 31(10): 3270-3282.

[4] Liu F L, Zhou H J, Gu B Y, et al. Application of large language models in medicine [J]. Nat Rev Bioeng, 2025, 3 (6): 445-464.

[5] Bedi S, Liu Y T, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review[J]. JAMA, 2025, 333(4): 319-328.

[6] 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.

[7] 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.

[8] Everett S S, Bunning B J, Jain P, et al. From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis[J]. NPJ Digit Med, 2026, 9: 409.

[9] Jin Q, Wang Z F, Floudas C S, et al. Matching patients to clinical trials with large language models [J]. Nat Commun, 2024, 15(1): 9074.

[10] Omiye J A, Gui H W, Rezaei S J, et al. Large language models in medicine: the potentials and pitfalls: a narrative review[J]. Ann Intern Med, 2024, 177(2): 210-220.

[11] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance [M]. Geneva: World Health Organization, 2021.

[12] World Health Organization. Regulatory considerations on artificial intelligence for health [M]. Geneva: World Health Organization, 2023.

[13] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models [M]. Geneva: World Health Organization, 2024.

[14] International Medical Device Regulators Forum. Software as a medical device (SaMD): clinical evaluation: IMDRF/SaMD

- WG/N41FINAL; 2017 [EB/OL]. (2017-09-21) [2026-10-12]. <https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation>.
- [15] U. S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions [EB/OL]. (2025-08-18) [2026-10-12]. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>.
- [16] U. S. Food and Drug Administration. Artificial intelligence-enabled device software functions: lifecycle management and marketing submission recommendations; draft guidance [EB/OL]. (2025-01-07) [2026-10-12]. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing>.
- [17] European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [J]. Official Journal of the European Union, 2024, L 2024/1689: 1-113.
- [18] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0): NIST AI 100-1 [R]. Gaithersburg: NIST, 2023.
- [19] Autio C, Schwartz R, Dunietz J, et al. Artificial intelligence risk management framework: generative artificial intelligence profile: NIST AI 600-1 [R]. Gaithersburg: NIST, 2024.
- [20] Vassilev A, Oprea A, Fordyce A, et al. Adversarial machine learning: a taxonomy and terminology of attacks and mitigations: NIST AI 100-2e2025 [R]. Gaithersburg: NIST, 2025. DOI: 10.6028/NIST.AI.100-2e2025.
- [21] International Organization for Standardization. ISO 13485:2016 medical devices—quality management systems—requirements for regulatory purposes [S]. Geneva: ISO, 2016.
- [22] American National Standards Institute, Association for the Advancement of Medical Instrumentation. ANSI/AAMI/ISO 14971:2019 medical devices—application of risk management to medical devices [S]. Arlington: AAMI, 2019.
- [23] American National Standards Institute, Association for the Advancement of Medical Instrumentation. ANSI/AAMI/IEC 62304:2006 and A1:2016 medical device software—software life cycle processes [S]. Arlington: AAMI, 2015.
- [24] International Organization for Standardization. ISO/IEC 23894:2023 information technology—artificial intelligence—guidance on risk management [S]. Geneva: ISO, 2023.
- [25] International Organization for Standardization. ISO/IEC 42001:2023 information technology—artificial intelligence—management system [S]. Geneva: ISO, 2023.
- [26] International Electrotechnical Commission. IEC 81001-5-1:2021 health software and health IT systems safety, effectiveness and security—part 5-1: security—activities in the product life cycle [S]. Geneva: IEC, 2021.
- [27] Center for Medical Device Evaluation, National Medical Products Administration. Guidance for registration review of artificial intelligence medical devices (Announcement No. 8 of 2022) [R]. Beijing: CMDE, 2022. (in Chinese)
- [28] Center for Medical Device Evaluation, National Medical Products Administration. Guidance for registration review of medical device software (2022 revision) (Announcement No. 9 of 2022) [R]. Beijing: CMDE, 2022. (in Chinese)
- [29] Center for Medical Device Evaluation, National Medical Products Administration. Guidance for registration review of medical device cybersecurity (2022 revision) (Announcement No. 7 of 2022) [R]. Beijing: CMDE, 2022. (in Chinese)
- [30] General Office of the National Health Commission of the People's Republic of China, General Office of the National Administration of Traditional Chinese Medicine, General Office of the National Disease Control and Prevention Administration. Reference guide to artificial intelligence application scenarios in the health sector (Guo Wei Ban Gui Hua Han [2024] No. 420) [R]. Beijing, 2024. (in Chinese)
- [31] Cyberspace Administration of China, National Development and Reform Commission, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, National Radio and Television Administration. Interim measures for the management of generative artificial intelligence services (Order No. 15) [R]. Beijing, 2023. (in Chinese)
- [32] Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, National Radio and Television Administration. Measures for labeling artificial intelligence-generated synthetic content (Guo Xin Ban Tong Zi [2025] No. 2) [R]. Beijing, 2025. (in Chinese)
- [33] Yang D W, Zhou J, Chen R C, et al. Expert consensus on the metaverse in medicine [J]. Clin eHealth, 2022, 5: 1-9.
- [34] 白春学. 元宇宙医学之我见 [J]. 中国医药导刊, 2023, 25(1): 1-6.
- [35] 白春学. 医学GPT研发应用之吾见 [J]. 元宇宙医学, 2025, 2(1): 6-14.
- [36] 白春学. 肺结节专家——BAIMGPT白皮书 [J]. 元宇宙医学, 2025, 2(2): 55-64.
- [37] 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 footprint, social benefit, and economic value [J]. Metaverse Med, 2026, 3(1): 1-10.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs) [J]. NPJ Digit Med, 2024, 7(1): 183.
 - [42] Miao B Y, Chen I Y, Williams C Y K, et al. The MI-CLAIM-GEN checklist for generative artificial intelligence in health[J]. Nat Med, 2025, 31(5): 1394–1398.
 - [43] Tam T Y C, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review[J]. NPJ Digit Med, 2024, 7(1): 258.
 - [44] 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.
 - [45] 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.
 - [46] 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.
 - [47] 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.
 - [48] 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.
 - [49] 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.
 - [50] Tejani A S, Klontzas M E, Gatti A A, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update [J]. Radiol Artif Intell, 2024, 6(4): e240300.
 - [51] Elvidge J, Hawksworth C, Aşar 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.
 - [52] Chen Y L, Yang K H, Marušić A, et al. A reporting tool for practice guidelines in health care: the RIGHT statement [J]. Ann Intern Med, 2017, 166(2): 128–132.
 - [53] Guyatt G, Vandvik P O, Iorio A, et al. Core GRADE 7: principles for moving from evidence to recommendations and decisions[J]. BMJ, 2025, 389: e083867.
 - [54] Alber D A, Yang Z H, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks [J]. Nat Med, 2025, 31(2): 618–626.
 - [55] Yang Y F, Jin Q, Huang F R, et al. Adversarial prompt and fine-tuning attacks threaten medical large language models [J]. Nat Commun, 2025, 16(1): 9011.
 - [56] Zhang H, Huang J, Mei K, et al. Agent security bench: formalizing and benchmarking attacks and defenses in LLM-based agents [C]//Proceedings of the 13th International Conference on Learning Representations (ICLR): 2025.
 - [57] Goh E, Bunning B, Khoong E C, et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance [J]. Commun Med, 2025, 5: 59.
 - [58] 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.
 - [59] Olson K D, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout [J]. JAMA Netw Open, 2025, 8(10): e2534976.
 - [60] van de Sande D, Chung E F F, Oosterhoff J, et al. To warrant clinical adoption, AI models require a multi-faceted implementation evaluation [J]. NPJ Digit Med, 2024, 7(1): 58.
 - [61] Pearlman K, Wan W, Shah S, et al. Use of an AI scribe and electronic health record efficiency [J]. JAMA Netw Open, 2025, 8(10): e2537000.
 - [62] Dennstädt F, Hastings J, Putora P M, et al. Implementing large language models in healthcare while balancing control, collaboration, costs and security[J]. NPJ Digit Med, 2025, 8(1): 143.
 - [63] 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.
 - [64] Li H Y, Fu J F, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline[J]. J Med Internet Res, 2025, 27: e71916.
 - [65] Liu M X, Ning Y L, Teixayavong S, et al. A scoping review and evidence gap analysis of clinical AI fairness[J]. NPJ Digit Med, 2025, 8(1): 360.
 - [66] Omiye J A, Lester J C, Spichak S, et al. Large language models propagate race-based medicine [J]. NPJ Digit Med, 2023, 6(1): 195.

引用本文

宋元林,高承实,杨达伟,等.可信医学智能体:研发、评价与临床实施中国专家共识[J].元宇宙医学,2026,3(3): 188–199.

Song Y L, Gao C S, Yang D W, et al. Trustworthy medical agents: a Chinese expert consensus on development, evaluation, and clinical implementation[J]. Metaverse Med, 2026, 3(3): 188–199.