

DOI: 10.61189/330547lpbnmt

· 技术应用 ·

科研智能体能否产生可重复的净科研增益? TRACE-Med 临床研究框架

李和权¹, 季婷芬¹, 白春学^{2,3,4,5*}

1. 浙江大学医学院附属邵逸夫医院呼吸内科, 杭州 310009
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] 医学研究面临的挑战不仅是信息过载, 更是证据、问题、研究、实施与复现之间的断裂。科研智能体可通过调用外部工具、整合多源信息并执行多步骤任务, 但同样可能把虚假引文、确认偏倚、版本漂移和越权操作放大为系统性科研错误。现有医疗人工智能评价多集中于模拟任务, 严格定义的智能体仍缺乏系统研究; 目前亦缺乏把科研智能体视为复杂研究干预的统一定义、可比较终点和前瞻性验证路径。本文据此提出可证伪命题: 科研智能体应在预先设定的研究质量标准下, 以提高可重复的净科研收益且不增加严重科研错误风险作为评价依据。TRACE-Med 将工作流组织为 Track(追踪证据)、Refine(凝练问题)、Architect(架构研究)、Coordinate(协同实施)和 Evaluate(验证复现)五阶段, 并以证据溯源、预先规定、权限分离和独立重跑作为假定作用机制。临床意义: 本文把最小可实施科研单元、有效且可重复的科研增益(VRRG)与智能体诱导科研错误率(AIRER)作为未来评价科研智能体实施效果的候选组件, 而非已获验证的量表或疗效终点。当前证据支持受控增强, 不支持无监督替代; 下一步应采用盲法结局判定、随机或交叉任务设计、多中心实施研究和独立复现, 检验框架是否产生净获益。

[关键词] 科研智能体; 临床研究; 可重复性; 人工智能治理; 开放科学; TRACE-Med

[中图分类号] R 4 **[文献标志码]** A

Can research agents deliver reproducible net research gain The TRACE-Med framework for clinical research

Li Hequan¹, Ji Tingfen¹, Bai Chunxue^{2,3,4,5*}

1. Department of Respiratory Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310009, Zhejiang, 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] The challenge facing medical research is not only information overload but also the disconnects among evidence, questions, research, implementation, and replication. Research agents can call external tools, integrate multi-source information, and perform multistep tasks, but they may likewise amplify fabricated citations, confirmation bias, version drift, and unauthorized actions into systemic research errors. Existing evaluations of medical artificial intelligence have mostly focused on simulated tasks, and rigorously defined agents remain understudied; there is also a lack of unified definitions, comparable endpoints, and prospective validation pathways for treating research agents as complex research interventions. Accordingly, this article proposes a falsifiable proposition: research agents should be evaluated under prespecified research quality standards by whether they increase reproducible

[收稿日期] 2026-07-18

[接受日期] 2026-09-12

[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] 李和权, 博士, 主任医师; E-mail: lihequan@zju.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

net research benefit without increasing the risk of serious research errors. TRACE-Med organizes the workflow into five stages—Track (evidence tracing), Refine (question refinement), Architect (research architecting), Coordinate (coordinated implementation), and Evaluate (validation and replication)—and uses evidence provenance, prespecification, separation of privileges, and independent reruns as putative mechanisms of action. Clinical implications: This article positions the minimum implementable research unit, Valid and Reproducible Research Gain (VRRG), and Agent-Induced Research Error Rate (AIRER) as candidate components for future evaluation of the implementation effects of research agents, rather than as validated scales or efficacy endpoints. Current evidence supports controlled augmentation, not unsupervised replacement; next steps should employ blinded outcome adjudication, randomized or crossover task designs, multicenter implementation studies, and independent replication to test whether the framework produces net benefit.

[Key Words] research agents; clinical research; reproducibility; artificial intelligence governance; open science; TRACE-Med

本文是批判性方法学综述,目标是界定科研智能体领域的核心科学问题并提出可前瞻性验证的研究框架,而非估计总体效应。证据基础包括已发表科学发现研究、系统或范围综述、报告规范与治理共识^[1-56]。由于初稿未提供数据库、完整检索式、检索日期、双人筛选流程或预注册方案,本文不能被表述为系统综述,也不主张实现文献检索的完全覆盖;所有定量结果均保留原始出处及其研究设计边界。涉及TRACE-Med、最小可实施科研单元和评价指标的内容均为作者提出的概念框架,尚未经过前瞻性验证。

1 临床问题:自动化能力增长快于可靠性证据

医学研究的关键损失往往发生在多个界面:问题选择与证据无关,可获得数据替代临床问题,方案与实施脱节,分析策略受到结果信息影响,数据与代码缺乏谱系,失败路径和阴性结果未被保存。选择性报告、低统计效能、分析自由度过大与不透明工作流可使统计显著性和可重复知识分离^[41,43-45]。科研智能体之所以值得研究,不是因为它能生成更多文本,而是因为它可能把证据检索、研究设计、实施质控、统计分析与版本管理整合成连续工作流;当然,这也能导致一次虚假引文或错误假设跨阶段传播。

本文把科研智能体定义为:以明确科学问题为目标,能够保持跨步骤任务状态,调用文献、数据库、统计软件或试验管理工具,在预设权限内依据反馈调整行动,并保留证据—决策—数据—代码审计链的人工智能系统。ReAct与Toolformer说明模型可把推理和外部工具调用结合^[3-4],但工具调用并不代表科学有效性。缺少持续目标、工具调用、版本状态、反馈迭代或全过程审计任一条件的系统,更适宜称为AI辅助工具。评价对象应是研究者—智能体—数据—工作流系统,而不是单一的基础模型。

2 知识缺口:现有研究证明了任务可行性,但尚未证明净获益

科学发现提供了科研智能体技术可行性的初步证据。Coscientist和ChemCrow可规划化学任务并调用实验或专业工具^[6-7];Virtual Lab由智能体协作设计92种SARS-CoV-2纳米抗体,并通过实验验证筛选具有功能潜力的候选物^[8]。这些研究支持一种有限机制解释:当模型输出与外部工具接地、由可量化反馈更新并接受现实世界验证时,智能体可扩大候选空间和协调复杂任务^[1-2]。它们并未证明相同系统能够改善临床研究的偏倚控制、受试者安全或独立复现。

临床人工智能的证据累积速度明显滞后于发表数量增长。一项覆盖2022年至2025年9月的LLM辅助系统综述纳入4 609项同行评议研究,其中1 048项使用真实临床数据,只有19项为前瞻性随机试验^[9]。另一项医疗LLM评价综述纳入519项研究,约5%使用真实患者照护数据,公平性、偏倚和毒性评价不足^[10]。严格定义医疗智能体的范围综述截至2025年4月只纳入7项研究,其中仅1项为42例、三组随机试验;其余多为模拟、计算、可行性或诊断准确性研究^[5]。因此,“能完成任务”不能替代对真实工作流、患者相关结局和科研错误的评价。

人机协作也不是天然的增益机制。106项实验、370个效应量的预注册荟萃分析显示,人—AI组合平均弱于人或AI中表现最佳的一方(Hedges $g = -0.23$, 95% CI -0.39 至 -0.07)^[13]。这项跨任务结果不能直接量化医学科研智能体的效果,却足以反驳“加入人工审核必然安全”的假设。自动化偏倚、责任稀释、锚定和相关错误可能把审核变成形式。科研智能体当前合适的定位是受控增强:优先承担高频、可验证、可回滚任务;由人类保留问题优先级、因果判断、伦理安全和不可逆决定。

表 1 关键证据、定量结果及其解释边界

证据来源	原始定量结果	可支持与不可支持的结论
临床 LLM 证据图谱 ^[9]	4 609 项研究;1 048 项真实临床数据;19 项前瞻性随机试验	覆盖面广但依赖 LLM 辅助筛选/分层;不能直接推断科研智能体有效
医疗智能体范围综述 ^[5]	7 项研究;仅 1 项随机试验(42 例、3 组)	定义严格但样本小、异质性高;多数未进入真实 workflow
医疗 LLM 评价综述 ^[10]	519 项评价;约 5% 使用真实患者照护数据	提示真实性、公平性与毒性评价不足;研究对象不等于科研智能体
人—AI 协作荟萃分析 ^[13]	106 项实验、370 个效应量;相对最佳单方 Hedges $g=-0.23(95\% \text{ CI } -0.39 \text{ 至 } -0.07)$	跨任务平均效应;不能假设协作必然优于人或 AI 单独工作
Virtual Lab 实验 ^[8]	设计 92 种纳米抗体;实验获得功能性候选物	证明计算—实验闭环的可行性;不是临床研究有效性证据
TrialGPT 研究 ^[48]	183 个合成患者、>75 000 条标注;召回>90% 相关试验;小型用户研究筛选时间减少 42.6%	合成数据与小型用户研究;不支持对真实招募结局作因果外推

3 核心框架:TRACE-Med 将科研智能体定义为可审计干预

TRACE-Med 并非构建自动化写作流程,而是把每项关键主张绑定到来源、决策理由、数据对象、代码版本、责任人和可复现结果。五个阶段为 Track、Refine、Architect、Coordinate 和 Evaluate。任

何阶段均可因新证据、可行性失败或安全信号回退,但必须保留原版本和变更理由。三类职责分别回答“已知什么”(证据智能体)、“如何可靠地知道”(设计智能体)和“是否按计划得到并能再次得到”(试验与分析智能体);职责分离用于失败隔离,而不是以多个同质模型模拟独立审查。

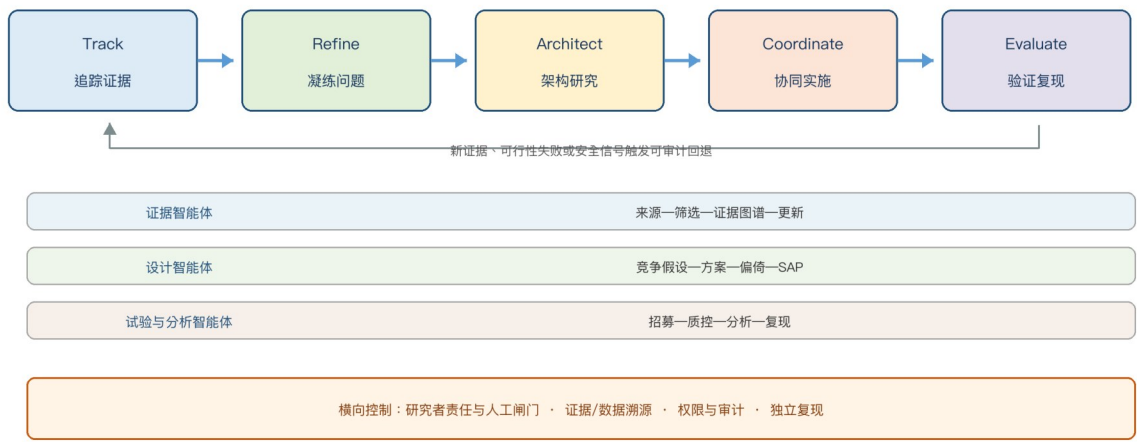


图 1 TRACE-Med 从证据发现到独立复现的闭环

该图为概念模型。回退须保留版本和理由;横向控制包括责任与人工闸门、证据/数据溯源、权限与审计以及独立复现。SAP=统计分析计划。

3.1 Track 与 Refine: 从信息召回到可证伪问题 Track 应建立可追踪的证据获取链,包括保存数据库、平台、检索式、日期、命中记录、去重规则、纳排理由和全文定位。PRISMA 2020 与 Cochrane 手册要求透明报告识别、筛选、偏倚评价和综合过程^[14-15];主动学习可减少低优先级筛选工作,但必须保存停

止规则、漏检估计和人工抽查^[16-19]。每个关键主张应连接唯一标识、研究人群、干预或暴露、结局、效应量、不确定性、偏倚域和适用边界,并分别使用 GRADE、AMSTAR 2、RoB 2 或 ROBINS-I 评价相应证据^[20-23]。Refine 随后建立竞争假设、反证检索、关键混杂、可观察预测和明确证伪条件。

表 2 TRACE-Med 各阶段的科学问题、输出和强制人工闸门

阶段	核心科学问题	智能体任务	最低可核查输出	人工闸门
Track	现有证据及其冲突/空白是什么?	可复现检索、去重、证据对象与更新监听	检索式、日期、纳排理由、效应量、偏倚域	信息专家与双人关键复核
Refine	问题是否重要、可检验且可证伪?	竞争假设、反证检索、混杂与可观察预测	替代解释、证伪条件、临床阈值	临床一患者一方法学共同确认
Architect	何种设计可识别目标效应?	估计目标、终点、样本量、随机/盲法、SAP与DMP	预注册方案、SAP、研究理由日志	研究者、统计与伦理批准;结果未知时冻结
Coordinate	研究是否按批准方案执行?	预筛、随访、质控、偏离与安全信号汇总	EDC查询、偏离/安全/版本日志	入排、改数、变更与安全决定由授权人员签署
Evaluate	结果是否稳健、可重跑、可外推?	锁库后分析、假设检查、敏感性分析与外部验证	可重复研究包、独立重跑与差异解释	独立代码审阅和责任作者确认

SAP:统计分析计划;DMP:数据管理计划;EDC:电子数据采集。

3.2 Architect 与 Coordinate: 将规范前移为执行约束 Architect 应先明确研究问题属于预测、因果还是决策任务,再规定估计目标、时间零点、终点、样本量、随机化、盲法、缺失数据与敏感性分析。SPIRIT 2025、CONSORT 2025 和 ICH E6(R3) 分别强化预先规定、开放科学和质量源于设计^[24-26];涉及 AI 干预时需按研究类型叠加 SPIRIT-AI、CONSORT-AI、DECIDE-AI、TRIPOD+AI、PROBAST+AI、STARD-AI、MI-CLAIM 或 REFORMS^[27-34]。Coordinate 可以生成招募预筛、随访、EDC 查询、方案偏离和安全信号提示。TrialGPT 在 183 个合成患者和超过 75 000 条试验标注中,以不足 6% 的候选试验召回超过 90% 的相关试验;小型用户研究中筛选时间减少 42.6%^[48]。这些结果支持辅助筛选的效率假设,但并不支持其自主决定受试者入排。已有关于 AI 辅助招募与研究保留的综述提示潜力,并受样本量、隐私、偏倚和结局不标准限制^[46]。

3.3 Evaluate: 将“可运行”提升为“可独立复现” 分析智能体只能在锁库后的隔离环境中执行正式代码,并检查分布、缺失、模型假设、数据泄漏、重复记录及结果一图表一正文一致性。主要结果至少应接受独立代码复核、从原始输入的非状态重跑和预设敏感性分析。可重复研究包必须包括数据字典、处理谱系、原始数据完整性校验信息、代码、依赖环境、随机种子、模型/提示/知识库版本以及人工修订和运行日志。因涉及隐私无法公开个体数据时,可采用受控访问、合成数据或可执行环境;“不可公开”不应等同于“不可验证”^[39-42]。

3.4 假定作用机制及其可证伪性 TRACE-Med 的机制解释是待检验假设,而非已证明的因果链。(1)

来源定位和稳定标识符可能减少虚假引文及限定条件丢失^[11-12]; (2) 预注册、研究理由日志和结果盲态可能减少结果后选择; (3) 最小权限、读写分离、双人授权和只追加日志可能阻断错误跨阶段传播; (4) 独立重跑与版本冻结可能提高计算可重复性。若这些中介指标改善而严重错误、人工负担或独立复现没有改善,则框架的机制假设不成立。

4 验证策略:以净科研增益和新增科研错误为双锚

最小可实施科研单元(MVRU)包括一个可证伪问题、时间戳方案、可追溯证据库、数据管理计划、SAP、风险分层人工闸门、模型/工具/人工修订日志和可重复研究包。MVRU 用于确保干预在比较前可被复制,而非规定必须使用某一模型。VRRG 和 AIRER 作为双锚:前者关注在质量阈值下产生的可重复净增益,后者关注智能体新增且足以改变设计、分析或结论的错误。二者目前尚无测量学验证,不应直接作为共同主要终点。

4.1 推荐研究设计 科研智能体的验证应遵循风险递进原则,未来研究可考虑采用以下分阶段验证路径: (1) 使用冻结基准集比较常规流程、通用 LLM、单智能体和多智能体,评价准确性、召回、校准、时间、成本和错误谱; (2) 采用随机对照或交叉任务研究,明确随机单位是研究者、研究任务还是研究中心,并对错误严重度与结果可复现性实施盲法判定; (3) 开展前瞻性多中心研究,评价使用率、延迟、方案偏离、亚组公平性、错误拦截和研究周期; (4) 追踪研究产出能否独立复现并产生临床或政策转化。模型、提示、知识库或接口更新应视为干预改变,现有临床 AI 随机试验范围不断扩大,但报告质量和患者结局证据仍有限^[47]。

表 3 双锚评价及解释性指标:操作定义与证据边界

指标	操作性定义	推荐测量	当前证据边界
VRRG	在预设质量阈值下,相对对照流程的方案、数据、依从、复现及转化净增益	先分别报告各分量;复合权重、失败规则与最小重要差异经验证后预先冻结	概念性双锚之一;目前不是已验证量表
AIRER	智能体新增、经盲法专家确认且若未拦截会改变设计、分析或结论的错误/机会数	按严重程度、来源、可拦截性和实际影响分类;报告率差、率比及区间	需预先定义“机会数”和严重错误判定规则
ETR	可准确回溯至原始来源及位置的关键主张比例	DOI/PMID 解析加全文定位抽查;按主张重要性分层	可追溯不等于高确定性,需并列偏倚/GRADE
HFS	预测、反证条件、替代解释和关键混杂的预设完整度	在获知结果前由盲法方法学家按量表评分	需建立内容效度、评者一致性和阈值
IRR	独立人员仅凭研究包重现主要数值、方向和推断的比例	无口头补充重跑并记录环境、耗时与差异	区分计算可重复、稳健性与新数据复现
HCB	为达到质量阈值所需的专家时间、严重程度加权修改与认知负荷	时间日志、修改审计、NASA-TLX 或等效工具	低人工时间若伴随漏错增加不构成优势

AIRER:智能体诱导科研错误率;API:应用程序接口;DMP:数据管理计划;EDC:电子数据采集;HER:电子病历;ETR:证据可追溯率;HCB:人工纠错负担;HFS:假设可证伪性评分;IRR:独立复现率;LLM:大语言模型;MVRU:最小可实施科研单元;RRL:研究理由日志;RRP:可重复研究包;SAP:统计分析计划;TRACE-Med:可追溯医学科研智能体循环;VRRG:有效且可重复的科研增益。

4.2 统计分析原则 样本量应依据严重 AIRER 的非劣效界值或预设 VRRG 分量的最小重要差异,并计入研究者、中心、疾病领域与任务难度的簇集。AIRER 应预先定义分母(任务、决策机会或智能体动作),采用率差、率比及 95% 置信区间;计数过度离散时可用负二项模型。连续时间和成本通常右偏,应报告中位数、分位数或经预设变换/稳健模型估计的差异。重复任务和多中心数据宜采用混合效应或分层模型。若建立 VRRG 复合分数,权重、失败规则、方向性、缺失数据处理和多重性控制必须在解盲前冻结,并同时报告全部分量,避免速度提升掩盖质量下降。

5 临床意义:受控增强而非自主替代

科研智能体最可能首先在可逆、可验证且标准明确的环节产生价值,例如检索更新、证据定位、一致性检查、数据查询、确定性重算和可重复研究包生成。高风险节点——问题优先级、因果解释、伦理审批、受试者入排、原始数据更正、揭盲、安全判断、方案变更与最终结论——应由具名人员承担。一个主张进入论文前,宜依次经过来源提交、推断适配、数字/代码核验和责任作者裁决;多个使用相同训练语料的智能体不构成独立复核。

治理对象是可行动系统。FUTURE-AI、WHO、ICMJE 和 WAME 分别强调公平、可追溯、稳健、隐私、独立审计、AI 使用披露和作者责任^[35-38]。最低

控制应包括项目级用途与风险预注册、最小权限、沙箱、读写分离、不可变原始数据、版本化提示/检索库/API、异常告警、双钥匙授权、事件追踪和独立复现。区域性元宇宙医学、医学 GPT、BAIMGPT、云端协同、医院—社区—家庭模式及肺结节智能体文献可作为应用场景和组织设计案例^[50-56],但主要属于观点、白皮书或早期实践,不能作为 TRACE-Med 有效性的比较证据。

实施可分为 3 个阶段:工具辅助、工作流协同和受控闭环。升级条件应是前一阶段达到预设的证据可追溯严重错误、独立复现和安全阈值,而不是使用时间或文本产量。低资源机构可采用小模型、规则引擎和人工审核实现相同控制目标;降低计算复杂度不能降低伦理、溯源和复现标准。

本综述有 6 项主要局限:(1)文献选择未按可复现系统综述流程记录,存在选择和时效偏倚;(2)科研智能体的定义未统一,不同研究对象高度异质;(3)支持性实例主要来自计算、模拟、化学领域或早期临床场景,外推临床研究质量和患者结局受限;(4)TRACE-Med、MVRU、VRRG、AIRER、ETR、HFS、IRR 和 HCB 均为概念性组件,尚缺乏内容效度、评价一致性、构念效度、响应性和阈值验证;(5)商业闭源模型的版本漂移、不可重复性及机构间管理差异可能削弱外部效度;(6)框架由倡议者提出,且部分应用例证为作者团队相关文献^[50-56],需要独立团队验证并透明披露利益冲突。

综上所述,科研智能体的临床研究价值不应以自主性、文本产量或速度判定,而应以一个可证伪问题判定:它能否在不增加严重科研错误和人工负担的前提下,提高可独立复现的净科研增益。现有证据证明了工具接地和人机协作的可行性,也显示真实临床验证稀少且协作可能产生性能损失。TRACE-Med 因此应被视为待检验的复杂干预框架。只有当证据有来源、方案先制定、权限可约束、错误可追踪、分析可重复且结论由研究者负责时,科研智能体才可能从效率工具转化为可信的医学研究基础设施。

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 李和权:素材提供、文章撰写;季婷芬: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] Wang H C, Fu T F, Du Y Q, et al. Scientific discovery in the age of artificial intelligence [J]. *Nature*, 2023, 620 (7972): 47–60.
- [2] Krenn M, Pollice R, Guo S Y, et al. On scientific understanding with artificial intelligence [J]. *Nat Rev Phys*, 2022, 4: 761–769.
- [3] Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models [C]//International Conference on Learning Representations, 2023, arXiv:2210.03629.
- [4] Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools [J]. *Advances in Neural Information Processing Systems*, 2023, arXiv: 2302.04761.
- [5] 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.
- [6] Boiko D A, MacKnight R, Kline B, et al. Autonomous chemical research with large language models [J]. *Nature*, 2023, 624(7992): 570–578.
- [7] M Bran A, Cox S, Schilter O, et al. Augmenting large language models with chemistry tools [J]. *Nat Mach Intell*, 2024, 6(5): 525–535.
- [8] Swanson K, Wu W, Bulaong N L, et al. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies [J]. *Nature*, 2025, 646: 716–723.
- [9] Chen S F, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine [J]. *Nat Med*, 2026, 32(3): 1152–1159.
- [10] 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.
- [11] Walters W H, Wilder E I. Fabrication and errors in the bibliographic citations generated by ChatGPT [J]. *Sci Rep*, 2023, 13(1): 14045.
- [12] Gao C A, Howard F M, Markov N S, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers [J]. *NPJ Digit Med*, 2023, 6(1): 75.
- [13] Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: a systematic review and meta-analysis [J]. *Nat Hum Behav*, 2024, 8: 2293–2303.
- [14] Page M J, McKenzie J E, Bossuyt P M, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews [J]. *Bmj*, 2021: n71.
- [15] Higgins J P T, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions* [M]. 2nd ed. Hoboken: Wiley, 2019.
- [16] Marshall I J, Wallace B C. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis [J]. *Syst Rev*, 2019, 8(1): 163.
- [17] van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews [J]. *Nat Mach Intell*, 2021, 3: 125–133.
- [18] Elliott J H, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence–practice gap [J]. *PLoS Med*, 2014, 11(2): e1001603.
- [19] Dennstädt F, Zink J, Putura P M, et al. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain [J]. *Syst Rev*, 2024, 13(1): 158.
- [20] Guyatt G H, Oxman A D, Kunz R, et al. What is “quality of evidence” and why is it important to clinicians? [J]. *Bmj*, 2008, 336(7651): 995–998.
- [21] Shea B J, Reeves B C, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both [J]. *BMJ*, 2017, 358: j4008.
- [22] Sterne J A C, Savović J, Page M J, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials [J]. *Bmj*, 2019: 14898.
- [23] Sterne J A, Hernán M A, Reeves B C, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions [J]. *BMJ*, 2016, 355: i4919.
- [24] Chan A W, Boutron I, Hopewell S, et al. SPIRIT 2025 statement: updated guideline for protocols of randomised trials [J]. *BMJ*, 2025, 389: e081477.
- [25] Hopewell S, Chan A W, Collins G S, et al. CONSORT 2025 statement: updated guideline for reporting randomized trials [J]. *Nat Med*, 2025, 31: 1776–1783.
- [26] International Council for Harmonisation. ICH Harmonised Guideline E6(R3): Good Clinical Practice [S]. 2025.
- [27] 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.

- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] Kapoor S, Cantrell E M, Peng K, et al. REFORMS: consensus-based recommendations for machine-learning-based science[J]. *Sci Adv*, 2024, 10(18): eadk3452.
- [35] 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.
- [36] World Health Organization. Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models[M]. Geneva: World Health Organization, 2024.
- [37] International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: Section V: Use of Artificial Intelligence in Publishing[EB/OL]. [2026-10-08]. <https://www.icmje.org/recommendations/>.
- [38] Zielinski C, Winker M A, Aggarwal R, et al. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications[J]. *Curr Med Res Opin*, 2024, 40(1): 11-13.
- [39] Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. *Sci Data*, 2016, 3: 160018.
- [40] Nosek B A, Alter G, Banks G C, et al. SCIENTIFIC STANDARDS. Promoting an open research culture[J]. *Science*, 2015, 348(6242): 1422-1425.
- [41] Munafò M R, Nosek B A, Bishop D V M, et al. A manifesto for reproducible science[J]. *Nat Hum Behav*, 2017, 1(1): 0021.
- [42] Sandve G K, Nekrutenko A, Taylor J, et al. Ten simple rules for reproducible computational research[J]. *PLoS Comput Biol*, 2013, 9(10): e1003285.
- [43] Ioannidis J P A. Why most published research findings are false[J]. *PLoS Med*, 2005, 2(8): e124.
- [44] Begley C G, Ellis L M. Raise standards for preclinical cancer research[J]. *Nature*, 2012, 483: 531-533.
- [45] Collaboration O S. Estimating the reproducibility of psychological science[J]. *Science*, 2015, 349(6251): aac4716.
- [46] Lu X R, Yang C, Liang L, et al. Artificial intelligence for optimizing recruitment and retention in clinical trials: a scoping review[J]. *J Am Med Inform Assoc*, 2024, 31(11): 2749-2759.
- [47] 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.
- [48] 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.
- [49] Zhang K, Yang X, Wang Y F, et al. Artificial intelligence in drug development[J]. *Nat Med*, 2025, 31(1): 45-59.
- [50] Yang D W, Zhou J, Chen R C, et al. Expert consensus on the metaverse in medicine[J]. *Clin eHealth*, 2022, 5: 1-9.
- [51] 白春学. 元宇宙医学之我见[J]. *中国医药导刊*, 2023, 25(1): 1-5.
- [52] 白春学. 医学GPT研发应用之吾见[J]. *元宇宙医学*, 2025, 2(1): 6-14.
- [53] 白春学. 肺结节专家——BAIMGPT白皮书[J]. *元宇宙医学*, 2025, 2(2): 55-64.
- [54] 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.
- [55] 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.
- [56] 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

李和权, 季婷芬, 白春学. 科研智能体能否产生可重复的净科研增益? TRACE-Med临床研究框架[J]. *元宇宙医学*, 2026, 3(3): 207-213.

Li H Q, Ji T F, Bai C X. Can research agents deliver reproducible net research gain The TRACE-Med framework for clinical research[J]. *Metaverse Med*, 2026, 3(3): 207-213.