

AGI 时代的医学、公共卫生与健康系统重新 对账

——临床证据、患者安全与人群健康边界

Reckoning with Medicine, Public Health, and Health Systems in the Age of AGI

Clinical Evidence, Patient Safety, and the Boundaries of Population Health

医学／公共卫生／健康系统独立外发专业深稿 V1.1

作者：子君赋 | 出品：文明跃迁研究组

证据截点：2026 年 8 月 26 日

摘要

人工智能进入医疗以后，一个最容易发生的跨层误判是：模型在测试中更准确，因此患者就会获得更好的健康结果。医学、临床流行病学、循证医学、患者安全、公共卫生与健康系统研究恰恰说明，中间仍隔着完整的现实链条：研究对象是否与真实患者相同，风险估计是否校准，诊断结果是否改变临床决策，治疗是否真正改善患者结局，错误是否可被发现与接管，少数群体是否承担不成比例的伤害，部署是否改变 workflows、转诊与资源占用，系统更新后证据是否仍成立，以及个体层面的局部收益能否在卫生系统与人群尺度保持净收益。本文因此把“模型准确率可以属于 AI；临床净收益、患者安全和人群健康不能因此自动属于 AI”作为专业重新对账的起点。

本文定位为 Perspective／Conceptual Review，不是临床指南、系统综述、药械监管意见或特定患者诊疗建议。文章继承医学和公共卫生已经形成的成熟证据纪律：诊断研究必须区分敏感度、特异度、校准、外部效度与临床后果；预测模型需要透明报告、风险偏倚与适用性评估；AI 进入真实 workflow 后需要早期临床评价、前瞻性试验、真实世界证据、持续监测和版本再验证；患者安全要求把错误放入完整系统链条；公共卫生则要求把个体诊断准确转化为人群监测、干预效果、可及性、公平与健康系统韧性的进一步证明。[1-16] 这些规范和方法不是为 AI 设置特殊禁区，而是防止技术性能越过医学本来就要求的证据桥梁。

截至 2026 年，现实证据已经同时给出正反两种重要压力。Epic 脓毒症模型在一个大型学术医疗中心的外部验证中仅取得 0.63 的住院层面 AUC，且产生明显警报负担，说明开发环境表现和广泛部署不能替代独立外部验证。[17] 一项 50 名医生的随机试验中，获得 GPT-4 辅助并未显著提高医生在模拟复杂病例上的诊断推理评分；探索性次级分析中，模型单独完成同一任务的评分更高，说明“模型更强”不等于“人机团队自然更强”。[18] 另一方面，瑞典 MASAI 随机筛查试验将 105,934 名女性随机分组（19 名后续排除，105,915 名进入分析），观察到 AI 支持乳腺筛查提高癌症检出率、减少约 44% 的阅片量且未显著提高假阳性率；2026 年进一步报告的间期癌结果满足非劣效性。[19-20] 德国 PRAIM 是一项前瞻性、多中心、真实世界、观察性非劣效实施研究，覆盖 12 个筛查点、超过 46 万名女性和 119 名放射科医师；AI 支持组检出率更高而召回率未恶化，但医生自愿选择是否使用 AI，因此选择偏倚与实施差异仍限制因果外推。[21] 这些结果共同说明：AI 医疗既不能被“实验室优秀、现实无益”的悲观叙事封死，也不能被“准确率更高、临床价值自动成立”的乐观叙事提前结算。

本文进一步把 AI 医疗重新放回公共卫生与健康系统。2025 年全球 UHC 监测报告估计，2023 年仍有约 46 亿人没有充分获得基本卫生服务，2022 年约 21 亿人因自付卫生支出面临财务困难；WHO 同时将 2030 年全球卫生与照护人员短缺预测上调至约 1110 万人。[25-26] 这意味着 AI 降低专业稀缺、改善可及性的机会极其真实，但机会只有在基础设施、语言、数字能力、转诊、治疗供给、支付、患者信任与卫生人员能力共同成立时，才会转化为人群健康。公共卫生领域，WHO 的 EIOS 2.0 已在 110 多个会员国和约 30 个组织／网络中使用，并加入 AI 功能；2026 年 WHO 还公开展示 AI 支持的社区倾听用于霍乱响应。[27-28] 这些都是现实采用事实，却不能自动证明更快监测已经导致更低发病、死亡或更正当的公共卫生干预。

本文最终主张：技术进入医疗以后，首先必须回答的不是系统多聪明，而是患者是否因此得到更好的真实结果；进入公共卫生以后，则必须继续回答人群健康是否改善、谁仍被排除、卫生系统是否更有韧性。健康是文明生活的重要条件，但不是人生全部意义。医学可以证明何种干预更安全有效，不能以疾病风险、寿命或健康评分定义人的完整价值；AI 也不能因为预测更准而取得治疗权、公共权力或意义支配权。医学专业的贡献不是为文明

方向背书，而是在最接近生命后果的位置把“什么真的有益、对谁有益、在什么条件下有益”证明清楚，并在证据停止处把问题交给法律、伦理、经济、政治、社会与文明方向重新判断。

关键词：AGI 时代；人工智能医疗；循证医学；临床有效性；患者安全；真实世界证据；公共卫生；健康公平；健康系统韧性

Abstract

Artificial intelligence in health care creates a recurrent inferential error: a model performs better on a technical benchmark, therefore patients must experience better health outcomes. Medicine, clinical epidemiology, evidence-based practice, patient safety, public health, and health-systems research show why this inference does not hold without additional evidence. Between model performance and patient benefit lie calibration, external validity, workflow integration, human-AI interaction, clinical decisions, treatments, adverse events, subgroup effects, resource use, follow-up, system resilience, and population-level consequences.

This paper is a Perspective/Conceptual Review rather than a clinical guideline, systematic review, regulatory opinion, or medical advice. It distinguishes established or verifiable facts and formal materials (A), unsettled mechanisms and competing explanations (B), cross-domain structural hypotheses (C), and explicit normative commitments (F). It argues that diagnostic accuracy, regulatory authorization, or model reliability cannot by themselves establish clinical net benefit in every deployment context; public-health surveillance performance cannot by itself establish intervention effectiveness; and average performance cannot erase subgroup harm. At the same time, emerging randomized and real-world evidence in screening demonstrates that well-designed AI-supported workflows can improve clinically relevant process and screening outcomes in defined settings.

The central thesis is: model accuracy may belong to AI evaluation, but clinical net benefit, patient safety, and population health require independent medical and public-health evidence. The appropriate endpoint of medical AI is therefore not “a smarter model” but a defensible chain from technical performance to real patient outcomes, equitable population benefit, and a resilient health system. Health is a major condition of a good civilization, but it is not a total score of human worth or meaning.

Keywords: AGI era; medical AI; evidence-based medicine; clinical effectiveness; patient safety; real-world evidence; public health; health equity; health-system resilience

研究定位、对象边界与论证纪律

本文是一篇医学／公共卫生／健康系统专业深稿。它独立承担临床有效性、临床净收益、诊断与治疗安全、外部效度、真实世界证据、患者异质性、临床责任接口、流行病学与公共卫生监测、干预效果、健康公平、卫生资源和健康系统韧性的专业事实责任。它不是生命科学稿的扩写：生命科学可以研究生命机制、疾病机制、个体化与生命连续性，但临床医学必须回答某种诊断、干预或工作流是否真的使患者获益，公共卫生必须回答干预是

否在真实人群尺度改善健康。原生命科学稿已经明确自身不是临床指南，并要求生命事实不得自动生成权利、人格或政治资格；这正是本稿独立成立的停止点之一。

本文同样不替 AI／计算机科学判断模型架构、基准性能、形式保证和系统安全，不替法学与伦理学判定法律责任、知情同意、隐私和权利效力，不替经济学决定价格、支付和资源分配，不替政治学决定公共卫生权力与紧急权限，也不替哲学或文明方向规定文明最终目的。医学负责把健康后果证明清楚；一旦问题进入权利、强制、分配与终极价值，必须重新交接。

表 1 A/B/C/F 论证身份与医学边界

身份	本文中的典型内容	能够支持什么	不能自动推出什么
A 较稳定事实/正式材料	临床试验、外部验证、真实世界研究、流行病学数据、患者安全报告、正式指南与监管材料	支持限定患者、场景、时间、版本与终点下的事实判断	不能自动推出所有人群净收益、法律责任或文明方向
B 前沿/竞争解释	长期人机协作、技能变化、持续学习模型、直接面向患者的 LLM、组织适应	比较竞争机制并形成进一步研究	不能写成已经收敛的普遍医学结论
C 跨域结构判断	模型→ workflow→行为→临床结局→资源→信任/可及→人群健康；共同模型→相关错误→系统风险	形成可反驳的跨域假说和压力测试	不能伪装成某一临床试验已经证明的完整文明链
F 规范承诺	患者安全不被准确率替代；人的基本地位不由疾病风险或健康评分决定；健康不是人生总分	公开说明本文为何要求真实结果、异质性与可退出	不能替代临床事实、现行法、公共授权或个人价值选择

“AGI 时代”不等待一个完成态 AGI

本文使用“AGI 时代”描述高级 AI、生成式 AI 和智能体系统已经开始进入医疗知识、记录、诊断支持、影像、筛查、患者沟通、公共卫生情报与卫生系统管理的现实条件。FDA 截至 2026 年仍持续更新 AI-enabled medical device 清单，清单中的设备已按各自适用路径满足相应上市前要求，但 FDA 同时明确该清单并不完备。[5] WHO 也已把 AI 实际应用概括到诊断、临床照护、药物开发、疾病监测、疫情响应和卫生系统管理。[1-2,29] 这足以使本文成立，不需要把未来 AGI、超级智能或机器主体性当作前提。

目录

- 导论 | 模型跨过诊室门槛以后，医学才真正开始
- 一 | 成熟专业遗产：医学、公共卫生与健康系统已经知道什么
- 二 | 研究对象：从模型输出到真实患者和人群后果
- 三 | 证据不能跳级：从技术性能到临床净收益
- 四 | 诊断：准确率为什么还不是诊断安全
- 五 | 治疗与临床决策：建议正确不等于患者获益
- 六 | 人机协作：模型更强为什么团队未必自然更强
- 七 | 正面证据：AI 在什么条件下已经改善真实临床流程
- 八 | 真实世界：外部效度、漂移、版本与持续监测
- 九 | 患者异质性与健康公平：平均性能不能覆盖谁受损
- 十 | 生成式 AI 与患者直接使用：新的健康入口与新的证据缺口
- 十一 | 公共卫生：更早发现不等于干预已经有效

十二 | 可及性与卫生资源：专业稀缺能否真正转化为普惠
十三 | 健康系统韧性：共同模型、相关错误与数字基础设施依赖
十四 | 临床责任接口：谁必须在什么节点重新进入
十五 | 医学与文明方向：健康重要，但不是人生总分
十六 | 跨学科交接：证据止步以后谁重新进入
十七 | 什么事能够改变本文：最强反驳与可反驳研究议程
结语 | 真正的终点不是更聪明的模型，而是更好的真实健康结果
续读与共同研究入口
附录 A | 来源—主张边界登记
附录 B | 医疗 AI 证据跃迁检查表
生成式人工智能使用声明
参考文献
版本说明

导论 | 模型跨过诊室门槛以后，医学才真正开始

AI 与医学之间最危险的误解，不是把一个表现普通的模型误写成“万能医生”，而是把一个在限定任务上真正优秀的模型，直接推成真实患者一定获益。前者容易识别，后者更难，因为推导的第一步可能完全正确：模型确实可以在某些影像、筛查、预测或知识任务中达到很高性能。真正断裂发生在下一步——它进入谁的工作流，改变了什么判断，触发什么检查或治疗，漏掉谁，额外召回多少人，占用什么资源，错误是否被发现，患者是否能够拒绝，系统更新后证据是否还成立。

医学已有一整套专门处理这种“从指标到后果”距离的方法。WHO 2023 年 AI 健康监管考虑把风险—收益、预期用途、外部验证、数据质量、持续学习、人类干预和全生命周期监测放在同一框架中；WHO 关于健康大模型的指导进一步指出，模型可能生成不准确、不完整或虚假信息，产生自动化偏误、技能退化、隐私与公平风险。[1-2] 这不是对 AI 的特殊不信任，而是医学本来就不允许一个中间指标替患者结局结账。

患者安全使这一点更加直接。WHO《Global Patient Safety Report 2024》估计，诊断错误约占全球可预防患者伤害的 16%，并指出历史采集、检查、检验选择与解释、沟通等多个环节都可能参与诊断伤害。[4] AI 进入这些环节以后，可以帮助发现错误，也可以增加新的错误来源。问题不能被压成“模型总体准确率是多少”，因为医疗伤害沿着完整行动链发生。

模型准确率可以属于 AI；临床净收益、患者安全和人群健康不能因此自动属于 AI。

这也是为什么本文必须独立于生命科学、AI 科学与伦理法学。生命科学告诉我们疾病与生命机制，AI 科学告诉我们系统能做什么，伦理法学告诉我们什么权利与责任必须被尊重；医学与公共卫生仍需独立证明：这样做以后，真实患者和真实人群究竟变好还是变坏。

一 | 成熟专业遗产：医学、公共卫生与健康系统已经知道什么

1.1 循证医学：证据围绕临床问题和患者结局，而不是技术新颖性

医学中的证据从来不是“越新越高”。随机试验、诊断准确性研究、队列、病例对照、真实世界数据、机制研究、经验与患者偏好回答不同问题。AI 介入以后，这一基本结构没有失效。SPIRIT-AI 与 CONSORT-AI 要求 AI 干预试验明确版本、输入输出、人机交互、错误分析与实施环境；DECIDE-AI 则专门处理从离线性能进入早期真实临床使用时的人因、工作流和可推广性；这些都是报告规范，不是替研究结果授予有效性。[10-13]

Topol 较早把 AI 与临床专业人员的互补视为“高性能医学”的潜在方向；Kelly 等则从临床影响实现条件出发，强调稳健评价、可推广性、工作流整合、监管与持续监测等障碍。[33-34] 本文继承的是这条专业遗产：AI 的临床价值必须在真实医疗系统中被证明，而不能由技术潜力或单次基准表现预支。

1.2 临床流行病学：区分区分能力、校准、外部效度和决策后果

TRIPOD+AI 要求临床预测模型透明报告开发和评估信息，PROBAST+AI 将模型开发与模型评价分开，并分别审查参与者／数据、预测因子、结局与分析中的质量、偏倚和适用性；STARD-AI 则针对 AI 诊断准确性研究增加数据集实践、AI 指标测试、算法偏倚、公平与泛化等报告要求。[8-10] 这些工具的共同作用不是建立一个“AI 医学证据总分”，而是让读者知道一项性能究竟在哪里成立。

区分能力很高也不等于风险估计可靠。Van Calster 等强调，预测模型即使具有良好 discrimination，也可能严重失校准，进而误导临床决策；外部验证和必要的模型更新因此不可缺少。[15] Vickers 与 Elkin 提出的 decision curve analysis 则进一步把阈值与临床后果纳入模型评估，提醒准确性指标并不直接回答“使用这个模型是否值得”。[16] 本文所称“临床净收益”采用更宽含义：患者相关获益扣除伤害、错误干预、机会成本和系统负担后的总体临床结果；它不等同于 decision curve analysis 中的技术性 net benefit 指标。

1.3 患者安全：错误是系统事件，不只是部件事件

WHO《Global Patient Safety Action Plan 2021-2030》将安全文化、组织可靠性、临床流程安全、患者和家庭参与、卫生人员教育、信息研究与协作纳入系统性行动。[3] Bates 等对 AI 与患者安全的范围综述也指出，AI 既可能减少诊断、用药和监测错误，也可能通过错误建议、自动化偏误、集成失败和新的工作流风险造成伤害。[31] 因此，患者安全不能只问“模型是否会错”，还要问错误怎样被防止、被发现、被放大或被修复。

1.4 公共卫生：研究对象从患者扩展到人群

公共卫生关心发病、死亡、暴露、传播、预防、筛查、监测、干预、公平与健康决定因素。一个算法可以提高某种个体预测，但公共卫生仍要问：目标人群是否真正被覆盖，假阳性与假阴性怎样分布，干预是否实施，资源是否跟得上，行为与信任怎样变化，最终发病和死亡是否改善。监测系统“更早发现信号”是重要中间结果，不等于公共卫生干预已经有效。

1.5 健康系统：服务可达、人员、基础设施与韧性决定技术能否成为健康

WHO 2025 年 UHC 监测显示，2023 年全球仍约有 46 亿人不能充分获得基本卫生服务；2022 年约 21 亿人因卫生支出面临财务困难。[25] WHO 执行委员会 2025 年卫生人力材料又把 2030 年全球卫生与照护人员短缺预

测上调至约 1110 万人。[26] 这些不是“AI 一定能够解决”的证据，而是解释为什么降低专业稀缺、文书负担与知识门槛具有巨大潜在价值。技术若只改善高资源机构的局部效率，却不能增加真实服务可达，健康系统层面的文明红利仍未发生。

二 | 研究对象：从模型输出到真实患者和人群后果

“医疗 AI”不是单一对象。一个模型权重、一项 FDA 授权医疗器械、一个医院部署服务、一套医生 workflow、一个患者直接使用的聊天机器人、一个国家级监测平台，在证据对象上完全不同。FDA 的 AI-enabled medical device 清单只说明特定设备按相应路径获得美国市场授权，并不构成所有部署环境或患者群体的临床净收益证明。[5] 同样，一个未作为医疗器械监管的通用大模型可以影响患者行为，却不因此拥有与医疗器械相同的证据状态。

表 2 医疗 AI 的六层对象与证据停止点

对象层	真正研究什么	主要证据	不能自动推出
模型/算法	在定义输入与数据集上产生预测、分类或生成	性能、校准、鲁棒性、偏倚	真实临床净收益、治疗权限
AI 医疗产品/设备	带有预期用途、版本与功能的产品	监管资料、验证、风险管理	所有机构和人群均有效
部署服务	具体医院/平台中的端到端配置	本地验证、服务监测、故障与版本	组织 workflow 自然安全
临床 workflow	医生、护士、患者与 AI 共同形成决定链	人因、接管、时间、培训、责任接口	法律责任与权利最终裁决
患者路径	诊断、检查、治疗、随访和真实结局	临床试验、队列、真实世界证据	人群尺度自动有效
健康系统/人群	资源、可及、公平、监测、韧性、信任	流行病学、实施科学、卫生服务与公共卫生研究	文明最终价值排序

2.1 完整行动链

一个最小临床行动链至少包括：谁提出临床问题；采集什么病史、体征、影像和检验；AI 接收什么输入；输出是概率、分类、文本建议还是排序；谁看见输出；是否改变诊断、检查、处方、转诊或治疗；患者是否理解 AI 参与；谁能推翻系统；错误怎样被发现；实际治疗是否执行；患者结局怎样；新结果是否反馈进未来模型。任何一环变化都可能让“同一个模型”的临床效果改变。

2.2 监管授权、技术性能、临床效益与公共卫生效果必须分账

FDA 授权可以证明特定产品满足适用的上市前监管要求；技术评测可以证明特定版本在特定数据上表现；临床研究可以证明某种部署对患者结局或 workflow 的效果；公共卫生研究继续观察人群覆盖、资源、健康公平和群体结局。四者互相支持，却不存在自动转换。把它们混在一起，是医疗 AI 最常见的证据越权之一。

三 | 证据不能跳级：从技术性能到临床净收益

医疗 AI 需要的不是一条“证据金字塔”自动给模型评级，而是一条因果链。不同技术在不同用途下可能需要不同研究设计，但至少必须回答：模型是不是对正确的人、正确的输入、正确的结局作出可靠预测；这些预测是否

真的改变临床行为；改变是否改善患者相关结局；部署的错误与负担是否抵消收益；效果是否能跨机构、群体和时间维持。

表 3 从模型性能到临床与人群结果的证据跃迁

证据阶段	核心问题	典型证据	停止点
离线/回顾性性能	模型是否在定义数据上区分或预测	AUC、敏感度/特异度、误差、校准	不能证明临床使用有效
外部验证	换机构/时间/群体后性能是否保持	独立队列、校准、亚组表现	不能证明工作流与人因安全
早期真实临床评价	进入真实流程后怎样被使用	DECIDE-AI 类研究、人因、接管、错误	不能替大规模比较结局
比较性临床研究	AI 流程相对标准照护是否更好	随机试验、准实验、患者/临床结局	不自动证明长期/跨地区泛化
真实世界证据	规模部署后效果、罕见风险与异质性	EHR、注册、理赔、设备数据、监测	观察性关联需控制混杂
生命周期监测	版本、数据、实践变化后证据是否仍成立	漂移、事件、再校准、版本复核	不能用“曾经有效”永久授权

FDA 2025 年更新的医疗器械真实世界证据指南明确，其核心是评估真实世界数据是否具有足够质量生成可用于监管决策的 RWE，并未改变产品本身适用的证据标准。[6] 这说明真实世界证据并不是“低质量数据的替代品”，也不是只要数据量大就能绕过因果和偏倚。NICE 数字健康证据标准框架同样把技术性能、经济影响、部署与生命周期要求分开，并在 2022 年纳入自适应 AI 与数据驱动技术。[30]

医疗 AI 真正的证据升级，不是“模型更大”，而是主张距离真实患者后果越近，所需证据越必须重新进入。

四 | 诊断：准确率为什么还不是诊断安全

4.1 诊断不是一次分类

美国国家科学院 2015 年《Improving Diagnosis in Health Care》把诊断描述为随时间展开、涉及患者与专业人员合作、信息采集与解释的过程。[32] WHO 2024 年患者安全报告同样把历史采集、体格检查、检验选择和解释、沟通与随访纳入诊断错误路径。[4] 因此，AI 即使准确识别一张影像，也未必完成诊断；AI 即使正确给出鉴别诊断，也未必保证患者能得到及时检查和治疗。

4.2 外部验证失败可以让“已部署模型”重新成为研究对象

Wong 等对 Epic Sepsis Model 进行独立外部验证，在 27,697 名患者、38,455 次住院中观察到住院层面 AUC 为 0.63，并指出在研究阈值下存在低敏感度与显著警报负担；研究发生于单一学术医疗中心，不能外推所有医院，却证明“开发者报告表现+广泛部署”不能替代本地独立验证。[17] 后续不同医院结果出现差异，本身进一步说明场景、基础患病率、工作流和实施方式会改变模型价值。

4.3 需要同时看区分、校准、阈值与后果

一个风险模型可以把高危患者排在前面，却系统性高估或低估实际风险；一个诊断模型可以提高敏感度，却带来大量不必要检查；一个筛查系统可以提高病变检出，却主要增加低临床意义病变。STARD-AI、

TRIPOD+AI、PROBAST+AI、校准研究与 decision curve analysis 分别从报告、偏倚／适用性、风险估计可靠性与决策后果对这些问题提供方法。[8-10,15-16] 它们不能互相替代。

五 | 治疗与临床决策：建议正确不等于患者获益

从诊断进入治疗，证据门槛进一步提高。AI 可能推荐药物、剂量、检查、手术、转诊或随访，但建议本身不是治疗效果。药物和医疗干预的真实获益取决于患者特征、禁忌、共同疾病、相互作用、治疗依从、医疗资源和后续监测。模型对指南复述得越流畅，越容易制造“答案已经存在”的错觉，而临床决策恰恰经常发生在证据不完整、患者偏好冲突和风险需要权衡的地方。

SPIRIT-AI 与 CONSORT-AI 之所以要求 AI 干预试验描述输入、输出、交互、错误和实施环境，DECIDE-AI 之所以强调真实工作流与人因，根本原因是 AI 不是脱离临床系统的药片，而是复杂干预的一部分。[10-13] FUTURE-AI 2025 国际共识则把公平、普适性、可追溯、可用、稳健和可解释等原则转化为全生命周期实践建议；它是一份多方共识指南，不是证明某个 AI 产品有效的临床数据。[14]

5.1 “临床净收益”不是单一指标

本文不建立一个新的“临床净收益总分”。临床净收益是一个专业问题结构：与标准照护相比，患者相关获益是否真实增加；额外伤害、过度诊断、漏诊、不必要治疗、延迟、警报疲劳和机会成本是否可接受；不同患者群体是否得到相近或可解释的结果；这种收益能否维持。具体研究仍应选择适合的患者结局和分析方法。

5.2 高风险行动权必须和证据范围分开

AI／计算机科学稿已经通过“医疗代理”压力测试指出，诊断性能不能自动产生处方、取消预约、保险或资源配置权限。本稿从医学侧进一步补充：即使 AI 在建议层面已被证明具有临床价值，治疗权怎样配置、患者怎样同意、谁承担法律责任仍需法学、伦理和制度重新进入；医学只能说明何种专业复核和临床接管对安全是必要的，不能把专业权威偷换成公共权力。

六 | 人机协作：模型更强为什么团队未必自然更强

医疗 AI 真实部署常常是“人+AI”，因此单独比较模型与医生可能遗漏最重要的系统对象。Goh 等 2024 年随机临床试验纳入 50 名内科、家庭医学和急诊医生，在六个复杂临床病例上比较“医生+GPT-4+常规资源”和“医生+常规资源”。两组诊断推理评分没有显著差异（76% vs 74%；调整差异 2 个百分点，95%CI -4 至 8； $P=.60$ ）。在探索性次级分析中，模型单独完成同一病例的评分较常规资源组高 16 个百分点（95%CI 2 至 30； $P=.03$ ）。[18] 后一比较不是主要随机比较，不能与主要结局视作同等证据；研究本身又使用模拟病例、样本较小，因此不能证明 LLM 在真实临床中无益。它真正支持的是：给医生接上更强模型，并不会自动把模型能力完整转换成人机团队能力。

6.1 人机协作至少需要四种证据

第一，专业人员能否理解何时该信、何时该怀疑；第二，AI 输出是否改变决策且方向正确；第三，工作流是否给人足够时间、权限和替代路径；第四，长期使用是否改变人的技能、注意和责任。WHO LMM 指导明确把自

动化偏误、技能退化、错误信息和人际关系变化列为风险。[2] 这些目前仍有大量 B 类问题，不能先验宣布“AI 一定让医生退化”，也不能仅凭短期效率证明长期专业能力不受影响。

6.2 文书自动化：减负是真实价值，但不是患者结局证据

Duggan 等 2025 年对环境式 AI 文书工具的早期临床使用观察到文书时间和下班后记录负担下降；Olson 等 2025 年多机构前后研究也报告使用环境式 AI 记录工具后临床人员倦怠比例下降。[35–36] 这类研究为“AI 可以释放临床时间”提供现实正向证据，但主要终点是文书负担、时间和专业人员体验，且部分设计非随机。它们不能直接证明诊断更准确、患者死亡更少或医患关系必然改善。

七 | 正面证据：AI 在什么条件下已经改善真实临床流程

7.1 MASAI：从随机筛查进入间期癌结局

医疗 AI 并不缺少真正正向证据。瑞典 MASAI 随机筛查研究将 105,934 名女性随机分配至 AI 支持流程或标准双阅片，19 名后续排除，105,915 名进入分析。2025 年报告中，AI 组癌症检出率为每 1000 人 6.4 例，对照组为 5.0 例；假阳性率未显著增加，同时阅片量减少 44.2%。[19] 2026 年进一步发表的间期癌研究显示，AI 组间期癌率为每 1000 人 1.55 例，对照组为 1.76 例，并达到预设非劣效性标准。[20] 这使证据从“更多检出”继续向筛查安全的重要下游结局推进。

即便如此，MASAI 仍不能证明所有国家、所有设备、所有筛查年龄或单独 AI 读片都同样成立，也不能单凭筛查准确性回答死亡率、长期过度诊断、成本效益和不同人群公平。高质量正向试验不是“问题结束”，而是把哪些事实已经成立说得更精确。

7.2 PRAIM：大规模真实世界实施带来另一类证据

德国 PRAIM 是一项前瞻性、多中心、真实世界、观察性非劣效实施研究，嵌入德国人口乳腺筛查项目。论文摘要报告研究期间 463,094 名女性接受筛查；结果部分用于主要研究人群描述的分析样本为 461,818 名女性、119 名放射科医生，分布于 12 个筛查点。AI 组癌症检出率为每 1000 人 6.7 例，对照组 5.7 例，AI 组召回率没有恶化。[21] 它的重要性在于规模、真实临床环境和多中心实施；但医生是否使用 AI 为自愿选择，研究并非随机分组，因此选择偏倚和实施差异需要保留。随机试验与真实世界研究不是谁取代谁，而是分别补上因果效应与外部实施证据。

7.3 结肠镜：更高检出也要继续问“检出的是什么”

Makar 等 2025 年纳入 28 项随机试验、23,861 名参与者的荟萃分析发现，AI 辅助结肠镜提高腺瘤检出率并降低腺瘤漏诊率；与此同时，新增检出主要来自微小病变，晚期腺瘤检出没有显著差异，非肿瘤性切除率反而上升。[37] 这是一条很重要的医学纪律：中间结局改善是真实进展，但越接近患者长期获益，越需要继续证明，而不能把“检测到更多”自动等同于“总体更健康”。

八 | 真实世界：外部效度、漂移、版本与持续监测

医疗系统会变化：疾病谱、流行病、检测方法、临床指南、人口结构、设备、EHR 字段、转诊逻辑和医生行为都在变化。AI 系统也会变化：模型、数据、提示、检索源、阈值、界面和供应商可能更新。一个在 2025 年成立的验证结果并不自动覆盖 2027 年的新版本。

8.1 真实世界证据不是“上线以后看看”

FDA 2025 年最终 RWE 指南把真实世界数据的相关性与可靠性作为能否生成监管可用证据的核心问题，并明确 RWE 可以来自 EHR、注册、理赔、设备和患者生成数据等多类来源。[6] 这要求研究者继续处理缺失、测量误差、选择、混杂、结局定义和可追溯性。大规模数据不能消灭偏倚，只会把偏倚放大得更稳定。

8.2 版本变化必须触发证据再定位

对可更新模型而言，变化不是抽象“性能升级”，而是证据对象改变。输入数据、模型、阈值、输出方式、警报逻辑、工作流或适用人群任何一项改变，都可能要求重新验证。FDA 的 AI 器械与临床决策支持监管实践已经把预期用途、透明度、软件功能与生命周期作为重要边界。[5,7] 本文不把美国监管路径写成全球统一标准，只把它作为成熟监管实践如何分离软件功能、医疗用途与证据状态的一个可核验例子。

8.3 真实世界监测必须允许“有效性下降”成为可见事实

如果医疗系统只记录 AI 被使用了多少次，却不记录错在哪里、哪个群体失效、警报是否被忽略、临床行为怎样变化，就无法形成真实学习系统。持续监测应当包含版本、患者结构、关键性能、临床结局、重大事件、覆盖与拒绝、人工接管和修复；但这些数据也必须服从隐私、最小必要与患者权利，不能为了监测把所有患者生活变成无限数据源。

九 | 患者异质性与健康公平：平均性能不能覆盖谁受损

9.1 平均准确率可以与结构性不公平同时成立

Obermeyer 等 2019 年分析一个广泛使用的人群健康管理算法，发现算法以医疗费用作为健康需要的代理，在相同风险分数下黑人患者实际上病情更重；若修正这一偏差，获得额外照护的黑人患者比例将从 17.7% 上升至 46.5%。[22] 关键问题并不只是“模型看见了种族”，而是一个看似有效的代理变量把既有医疗可及差异编码为健康需要。医学 AI 的公平因此从问题定义和结局选择就已经开始。

9.2 本地公平修正不保证跨地点仍公平

Yang 等 2024 年在放射、皮肤和眼科多个医学影像任务及六个全球胸片数据集集中研究公平与泛化，发现本地数据分布中有效的公平修正并不保证在新的外部测试环境中保持最优公平。[24] 这说明公平不能只在开发数据集上做一次差异指标后永久结算。患者异质性、设备、疾病谱、基础患病率和医疗系统差异都需要进入外部验证。

9.3 LLM 公平需要多维度 and 参与式评估

Pfohl 等 2024 年以 Med-PaLM 2 和 EquityMedQA 建立多维度、参与式的人类评估方法，显示对长文本医疗回答的公平伤害需要不同专业和生活经验的评估者共同显影；作者也明确指出，这套工具不足以整体证明部署后的健康结果公平。[23] 因而，模型公平评测是必要条件，不是人群公平结果。

平均性能提高，不会自动注销少数群体更高的错误率、延误、过度干预或服务排除。

十 | 生成式 AI 与患者直接使用：新的健康入口与新的证据缺口

生成式 AI 改变了医疗最传统的一项入口：患者不再必须先进入医院或搜索引擎，便可以获得高度连贯、个性化、持续对话式的健康解释。WHO 关于 LMM 的指导把患者直接使用列为五类主要应用之一，同时明确虚假、不准确、有偏或不完整信息、隐私、自动化偏误和操纵风险。[2] 这一变化的机会与风险都不能被传统“医疗器械有没有批准”完全覆盖，因为大量通用模型并不以受监管医疗器械身份进入日常生活。

10.1 健康信息不是诊断，也不是治疗

患者直接使用 AI 可以降低语言和知识门槛，帮助准备就医问题、理解术语、整理病史和识别需要就医的信号。但只要系统开始替患者完成风险分层、建议停药、延迟就医或改变治疗，它就进入更高后果层。相同输出在健康教育、分诊提示与治疗建议中的证据要求不同。

10.2 直接可达提高以后，专业关系反而更需要重新设计

AI 可以使健康知识随时可得，但患者真正需要的往往不是答案本身，而是身体检查、检验、治疗资源、长期照护、与家庭共同决策及责任承接。生成式 AI 如果成为第一入口，应当研究何时主动升级给临床人员、怎样表达不确定性、怎样避免以流畅语言掩盖证据不足，以及在弱势群体中是否真正减少而非扩大延迟和误导。

十一 | 公共卫生：更早发现不等于干预已经有效

11.1 AI 已经进入公共卫生情报

WHO 2025 年推出 EIOS 2.0，用于开放来源公共卫生威胁早期发现，并加入新的 AI 功能；发布时系统已在 110 多个会员国和约 30 个组织／网络中使用，WHO 的 EIOS 现行官方页面进一步记录：到 2025 年底，已有 120 个国家和 30 个组织／网络使用该系统。[27,38] 这证明 AI 不是假设性公共卫生工具，而已进入现实全球卫生情报基础设施；但采用规模本身不能证明公共卫生结局已经改善。

11.2 发现信号以后，仍然要经历核验、解释与干预

EIOS 的价值在于帮助专家从网站、社交媒体等开放来源中更快识别相关内容，随后仍需验证与评估。[27] 公共卫生错误可以来自漏掉真实威胁，也可以来自把噪声误写成疫情、误读地域与群体信号、过度干预或破坏信任。AI 提高信号检出速度，不等于流行病学因果已经成立，更不等于采取某种强制措施已经正当。

11.3 AI 支持社区倾听：从“信息更多”走向“行动是否更准”

WHO 2026 年 EPI-WIN 霍乱活动公开展示 AI 支持的社区倾听：分析热线、社交媒体、广播、调查与一线报告，以发现疫情报告、担忧、谣言、服务缺口和就医障碍，并用于 WASH、风险沟通、疫苗接种和快速响应的定向决策。[28] 这是现实应用路径的 A 类材料，但活动与案例并不构成随机效果证明。真正公共卫生问题仍是：行动是否因此更早、更准、更公平，发病、死亡和服务中断是否下降。

十二 | 可及性与卫生资源：专业稀缺能否真正转化为普惠

AI 医疗最值得期待的一面，是把知识和专业服务的边际成本降下来。全球仍有巨大的服务缺口和卫生人员短缺，这使远程支持、自动文书、影像辅助、培训和患者教育具有现实意义。[25-26,29] 但“一个模型能回答”与“一个人真正获得医疗”之间还有设备、网络、语言、支付、检查、药物、手术、护理、转诊和交通等多个条件。

12.1 降低一次判断成本，不等于系统总成本下降

AI 若提高筛查敏感度，可能带来更多后续检查；若降低文书负担，节省的时间可能被转化为更多患者，也可能被新的生产率要求重新占满；若让患者更容易获得健康建议，可能减少低价值就诊，也可能增加不必要咨询。Makar 等结肠镜荟萃分析中非肿瘤性切除增加，就是“检测更强”可能同时增加资源使用的现实例子。[37] 所以资源节省必须由完整卫生服务链证明，不能由模型推断。

12.2 普惠必须观察谁仍然进不来

健康公平不仅是模型不同群体准确率，还包括基础数字接入、语言、残障可达、支付、地理、卫生人力、信任和救济。WHO 的全球卫生战略把健康公平、初级卫生保健、UHC 与韧性放在同一战略中，并把负责任 AI、数字健康系统和网络安全作为未来能力的一部分。[29] 这是方向与治理基线，不是某个 AI 系统已改善 UHC 的证据。

十三 | 健康系统韧性：共同模型、相关错误与数字基础设施依赖

当一个医院只有一个 AI 工具时，错误可能是局部事件；当大量医院共享同一模型、云平台、影像算法、身份接口和供应商，错误可能变成相关性事件。医疗系统本来通过专业冗余、复核、转诊和多种技术路径降低单点失败；AI 标准化可以提高质量和可达，也可能降低认知与基础设施冗余。

13.1 相关错误比单个错误更值得系统关注

一个模型偏差若同时作用于数百万患者，传统“偶发个体误诊”的风险结构会改变。共同模型也可能让不同机构在同一亚组、同一输入异常或同一更新版本上同时失败。因此，健康系统层面的安全评估应增加故障半径、共同依赖、备用流程、人工接管和跨机构事件共享，而不是只看单设备平均故障率。

13.2 网络安全与可退出属于健康系统证据

WHO 2023 监管考虑和 2025-2028 全球卫生战略都把网络安全、数据治理和韧性纳入数字健康与 AI 的现实边界。[1,29] 如果医院在深度集成后无法在 AI 服务中断时继续诊疗，技术性能再高也可能形成新的脆弱性。可退

出在医疗中不意味着“患者自己关掉按钮”就结束，而是拒绝或故障之后仍有可用人工路径、核心病历可访问、处方和转诊可继续、危急服务不被中断。

13.3 韧性不是要求所有系统永久双套运行

完整冗余具有真实成本，不是每个低风险功能都需要同等备份。合理做法应按后果、替代难度、恢复时间和依赖规模分级。AI 文书工具的临时故障与 ICU 预警或国家公共卫生监测中断显然不是同一风险。本文只要求韧性进入证据与设计，不把“越多冗余越好”变成绝对规则。

十四 | 临床责任接口：谁必须在什么节点重新进入

医学中的责任首先是一种专业实践关系：谁取得病史、谁解释结果、谁决定进一步检查、谁开立治疗、谁监测不良反应、谁与患者沟通、谁在病情恶化时重新评估。AI 加入后，责任链增加了模型开发者、产品制造者、医院采购、IT 与数据团队，但“节点更多”不能让临床后果无人承担。

表 4 医疗 AI 完整责任链中的专业交接

节点	医学首先要回答什么	主要责任进入者	进一步交接
数据/输入	信息是否足够、来源和质量是否适用	临床团队+数据/检验专业	隐私与数据合法性→法学
模型输出	性能、校准、不确定性、亚组限制	AI/计算机+临床专业	系统保证→AI；证据强度→认识论
临床解释	是否符合病例、需否复核或追加检查	医生/相关专业人员	治疗权与职业规范→法学/专业监管
患者决定	风险、选择与替代是否被充分沟通	患者+临床人员	同意与代理→伦理／法学
治疗/执行	干预是否按适应证安全实施	临床组织与专业人员	资源与支付→经济/制度
随访/修复	错误是否被发现、纠正、补救、学习	医疗机构+患者安全系统	赔偿/法律责任→法学
人群监测	效果和伤害是否在群体尺度变化	公共卫生/卫生服务研究	强制与公共权力→政治/法学

这张表不分配法律责任比例，也不把医生固定为所有 AI 错误的最终承担者。它只阻断一种责任逃逸：开发者说“只是工具”，医院说“已采购合规产品”，医生说“系统建议如此”，最终患者承受完整后果却无人能回答为什么。法律归责必须由法学重新进入，但医疗系统首先要保证事实链和专业决定链可被重建。

十五 | 医学与文明方向：健康重要，但不是人生总分

文明方向稿把“生存无忧→生活幸福→生命更有意义”作为递进目标。医学最直接支撑的是生命安全、基础健康与减少可避免痛苦，也通过健康、能力与照护条件进入生活幸福。但医学不能把自己的专业目标膨胀成文明总目标。一个人有疾病、残障、衰老或较短预期寿命，不因此拥有较低人格价值、较少政治资格或较少生命意义。

15.1 健康优化不能成为全社会单一优化函数

AI 可以预测风险、提醒运动、优化治疗和公共卫生资源，但如果“健康风险最低”被扩张成对所有生活方式、关系、职业与自由的统一管理尺度，医学指标就超过了自己的对象。医学事实可以说明风险，不自动取得替人选择怎样生活的权力。

15.2 照护不能被效率完全吞没

医疗的价值不仅是生物指标改善，还包括解释、安慰、共同决策、陪伴、照护和面对不确定与死亡。AI 可以帮助医生腾出文书时间、降低信息负担和扩大专业能力；如果这些效率最终换来更厚的临床关系，它就是人的能力扩展。若效率只是把每名医生安排更多患者、把人退到末端确认接口，技术进步可能与关系变薄同时发生。现有环境式 AI 文书研究提供了前者的现实可能，却仍需要患者体验与长期临床结果继续验证。[35-36]

健康是文明生活的重要条件，但不是人生全部意义。

十六 | 跨学科交接：证据止步以后谁重新进入

表 5 医学／公共卫生稿与相邻专业的严格交接

交接学科	首先能够回答什么	医学稿在哪里停止
生命科学	疾病机制、生命过程、生物个体化	不能替临床有效性、风险—收益、人群健康
AI／计算机科学	模型性能、系统保证、版本、权限与技术退出	不能替患者净收益、公共授权
认识论／科学哲学	证据、因果、模型与不确定性	不能替具体疾病/干预生产事实
认知科学	自动化偏误、判断、学习与人机协作机制	不能替临床结局与责任效力
伦理／法学／政治哲学	同意、隐私、权利、责任、人格与规范边界	不能替临床风险—收益数据
政治学	公共卫生权力、国家卫生能力、紧急治理	不能替流行病学效果与临床事实
经济学／政治经济	支付、成本、所有权、卫生资源分配	不能替治疗有效性与患者安全
社会学／人类学	医患关系、信任、疾病经验、照护文化与不平等	不能替临床试验或流行病学
系统科学	级联、韧性、相关失败与复杂系统结构	不能替具体医学事实
教育学	医学教育、能力形成、AI 时代专业训练	不能替患者临床结果
国际关系／战略安全	跨境卫生安全与国际危机中的战略互动	不能替疫情事实、国际法或临床判断
生态／地球系统	环境与健康共同物质条件、基础设施承载	不能用生态事实直接定义治疗与健康权利

16.1 真实因果继续向外，但证据标准不融合

AI 辅助诊断可能改变患者流量，进而改变卫生资源；资源变化会影响可及和不平等；公共卫生数据又会影响政府行动；健康系统对少数云和模型的依赖可能进入国家安全。现实因果可以一路向外，但每一步都必须由相应专业重新举证。医学不能因为最接近生命后果而取得对全部文明问题的上位权力。

十七 | 什么事实能够改变本文：最强反驳与可反驳研究议程

17.1 反驳一：AI 最终会比医生全面更强，所以这些中间证据门只是过渡

如果未来 AI 在真实开放环境、跨疾病、跨人群、跨机构和长时间尺度上持续优于当前医疗系统，本文当然需要重写人机分工。但“系统更强”仍不能取消患者异质性、治疗后果、权利、资源与公共卫生外部性。最可能改变的是人工复核的范围，而不是临床结果仍需被证明这一原则。

17.2 反驳二：RCT 太慢，AI 更新太快，传统医学证据已经不适用

部分正确。不是所有 AI 变化都需要重新做大型 RCT；真实世界证据、准实验、连续监测、分阶段评价和预先定义的更新计划可以更适应动态系统。[6,10-14] 但“传统 RCT 不适用”不能推出“技术性能可以替临床证据”。更新越快，越需要把哪个版本、哪个变化、哪个结局与哪种证据绑定。

17.3 反驳三：真实世界大规模采用本身就是最好的证据

采用能证明需求、可用性和组织选择，不能自动证明净临床收益。广泛使用的脓毒症模型仍可能在外部验证中表现差；大规模观察研究也可能受选择和混杂影响。[17,21] 真正强的现实证据需要比较、结局、时间和偏倚控制，而不是下载量或医院数量。

17.4 反驳四：AI 只做辅助，所以最终责任都由医生承担即可

如果医生拥有足够信息、时间、能力、权限和真实替代路径，临床责任当然可以由专业人员承担重要部分。但当系统输出被默认、时间被压缩、医生无法理解版本变化或机构不提供替代路径时，“医生签字”可能只是形式控制。责任应根据真实控制、可预见性、专业义务和制度安排重新分配，法律结论由法学决定。

17.5 反驳五：公平只要做群体指标达标即可

Obermeyer、Pfuhl 和 Yang 等工作共同表明，公平问题可以来自代理变量、问题定义、输出内容、数据分布与外部环境，并且本地修正可能在新环境失效。[22-24] 如果未来形成能够跨地区稳定预测部署公平结果的方法，本文应降低对逐场景再验证的强调；目前证据不支持一次性“公平认证”。

17.6 能够改变或取消本文的事实

- 若长期研究证明医疗 AI 的模型性能在绝大多数高后果场景都能稳定、无需额外临床证据地预测患者净收益，本文关于“性能—临床后果断桥”的主轴应显著收缩。
- 若通用医疗 AI 在跨机构、跨人群、跨时间与跨疾病中表现出稳定外部效度，并且部署变化对效果影响极小，本文对本地再验证与版本重审的要求应降低。
- 若大规模人机协作研究证明 AI 长期稳定提高临床判断、专业能力与患者结局而不产生显著自动化偏误或技能损失，相应 B 类风险判断必须缩小。
- 若公共卫生研究证明 AI 增强监测稳定转化为更低发病、死亡和更高公平，本文应加强正向人群健康判断。
- 若进一步专业预研证明临床与人群健康事实只是少量案例，且生命科学、AI 科学和法伦理能够在不越权的情况下完整承担，本文的独立成篇资格应重新审查甚至取消。

17.7 可直接进入的研究议程

- 跨机构、跨人群、跨时间外部验证：不仅报告 AUC，还报告校准、亚组、阈值和临床后果。
- 真实人机团队试验：比较模型单独、专业人员单独、不同人机界面和不同培训方式，而不是只比较模型与医生。
- 患者结局试验：把准确率、检出率、工作量继续连接到死亡、并发症、生活质量、过度诊断与长期随访。
- 版本与漂移研究：建立更新后何时需要再验证、何种变化可以桥接旧证据的经验规则。
- 患者直接使用生成式 AI：研究分诊升级、延迟就医、错误建议、健康素养、弱势群体与长期行为后果。
- 健康公平：把开发数据公平指标与真实可及、治疗、结局和救济连接，报告负结果。
- 公共卫生：比较 AI 辅助监测是否真正缩短验证—响应时间、改善干预命中并降低人群伤害。
- 健康系统韧性：开展共同供应商故障、网络攻击、模型错误相关化和人工退回路径的压力测试。

结语 | 真正的终点不是更聪明的模型，而是更好的真实健康结果

AGI 时代会不断把医学推向一个诱人的捷径：系统已经能完成过去只有专业人员才能完成的任务，所以医疗问题正在被技术解决。事实更复杂，也更有希望。AI 确实已经在部分筛查、检测、知识和文书场景中产生可测的正向结果；它也暴露出外部验证失败、人机协作转换不充分、群体偏差和系统依赖等新问题。两类事实必须同时被保存。

医学的专业价值正在这里显现：它不问机器像不像医生，而问真实患者有没有因此获益；公共卫生不问信号是不是更多，而问人群健康是否真的改善；健康系统不问数字化程度是不是更高，而问服务是否更可达、人员是否更能照护、系统是否更能在错误和危机中继续运行。

技术进入医疗以后，首先必须回答的不是系统多聪明，而是患者是否因此得到更好的真实结果。

当这个问题被证明清楚，文明才有资格继续问：这些新的健康能力，应当怎样被所有人真正拥有，又怎样不把人的生命重新压成一张风险表。

续读与共同研究入口

向上 | 如果继续追问：为什么临床事实最终会变成文明问题

进入《AGI 时代的学科重新对账——学科边界、相互作用与共同研究议程》V2.4，继续理解为什么学科必须保持证据独立，又为什么医疗、经济、教育、公共权力、关系与意义会在现实因果中重新连接。本文不要求读者先接受文明跃迁理论；任何临床反例都拥有反向修改整体判断的权利。

进入《AGI 时代人类文明跃迁的机遇、方向、方法与目标实现——为什么必须改变、可以走向哪里、怎样抵达》V1.1，继续讨论健康怎样进入“生存无忧—生活幸福—生命更有意义”的文明方向。该文是规范与方向性展开，不是本文临床证据的来源。

向下 | 如果继续追问：怎样把判断变成真实验证

优先进入临床试验、诊断研究、真实世界研究、患者安全、实施科学、流行病学、卫生服务和公共卫生现场研究。对于机器侧能力、权限、日志、停止和退出，再进入 AGI-COS（机器侧能力、权限、审计、停止与退出的工程治理候选框架）相关工程研究；对于高后果选择的多主体推演，可进入 Human-COS（高后果选择在不可逆之前进行多主体、多视角后果推演与判断支持的候选系统）等验证接口。它们只能作为待验证承载，不能反向证明医学事实。

向左 | 如果继续追问：谁必须重新进入

AI／计算机科学重新判断模型和系统保证；生命科学重新判断生命和疾病机制；认识论重新判断证据与外推；认知科学研究自动化偏误和人机判断；伦理法学处理同意、责任、隐私与权利；经济学处理支付、成本和资源；政治学处理公共卫生权力；社会学／人类学处理医患关系、信任与照护文化；教育学处理医学教育与专业能力形成。

向右 | 如果不同意本文：最值得进入哪些相反方向

应优先检验四类真正有力量的反方：第一，AI 在真实临床中可以比本文预期更快形成稳定患者净收益；第二，大规模真实世界证据足以替代传统阶段性评价；第三，人机协作问题主要来自界面与培训而非 AI 本身；第四，未来通用医疗 AI 可以使本地验证、专业接管和大量传统工作流成为不必要成本。只要新的数据能支持这些方向，本文应收缩，而不是把反例重新命名后吸收。

向前 | 什么专业工作可以直接改变本文

临床医生可以提交阴性试验、失效病例和跨机构复现；流行病学者可以建立亚组与外部效度证据；患者安全团队可以公开事故与近失事件；公共卫生研究者可以把“监测更快”继续追踪到干预与人群结局；卫生服务研究者可以测量 AI 对可及、排队、成本与人力的真实影响；患者和共同体可以指出专业研究遗漏的伤害、信任与可达问题。最有价值的参与，不是赞同，而是足以改变理论的专业工作。

文明未来链 | 如果这些问题被更好解决，人类可能生活在怎样的世界

《活的文明》三部曲负责把未来生活世界具体展开：当诊断、健康知识和照护能力更容易抵达普通人，医院、家庭、养老、社区、工作和公共生活可能怎样变化。它是可质疑、可重构的未来图景，不是本文的经验证据。星火（个体能力与关心延展的候选系统）、星云（文明协作与运行的候选承载）、AGI-COS 和 Human-COS 等继续承担不同现实承载想象，也必须接受临床事实、失败和患者经验的反向修订。

文明跃迁五卷书：<https://www.civitas.top/documents/p02-five-001/> | 公众总入

口：<https://www.civitas.top/documents/p01-entry-006/> | 完整文库：<https://www.civitas.top/library/>

附录 A | 来源—主张边界登记

表 A1 主要来源与主张边界

来源组	本文使用范围	不得推出
[1-2] WHO AI 健康监管/LMM 指导	AI 健康的正式国际考虑、应用类别与风险边界	不能证明某个模型有效或全球统一法律义务
[3-4] WHO 患者安全文件	患者安全与诊断伤害的系统事实	不能把所有诊断伤害归因于 AI
[5-7] FDA 医疗 AI/软件/RWE	美国监管授权、软件边界和真实世界证据实践	不能外推为全球法或所有部署净收益
[8-14] AI 医学报告/评价规范	透明报告、风险偏倚、早期临床评价与试验设计	规范本身不是有效性证据
[15-16] 校准/DCA	风险估计可靠性与临床决策后果方法	不构成完整患者结局证明
[17-18] 脓毒症/LLM 诊断试验	外部验证失败与人机团队转换问题的实证	不能证明所有医疗 AI 无益
[19-21] MASAI/PRAIM	定义筛查场景中的随机和真实世界正向证据	不能自动跨国家、设备、人群和终点泛化
[22-24] 公平研究	代理偏差、LLM 公平评估与跨场景公平泛化限制	不证明所有部署必然歧视
[25-29,38] WHO UHC/人力/公共卫生/全球战略	全球服务缺口、卫生人力、AI 公共卫生采用与健康系统方向	不能证明 AI 已改善 UHC 或疫情结局
[30-37] 健康技术评估/诊断/患者安全/文书/结肠镜	证据框架、成熟医学问题与多场景补充实证	受各自设计与终点限制

附录 B | 医疗 AI 证据跃迁检查表

表 B1 医疗 AI 从性能到真实结果的最低检查表

项目	必须回答	最低纪律
对象	当前说的是模型、产品、部署、工作流、患者路径还是人群系统？	避免把“AI”写成单一对象
预期用途	系统究竟辅助什么决定，谁使用，发生在什么时点？	用途变化必须重开证据
技术性能	区分、校准、鲁棒性和亚组表现是否充分？	不能停在 AUC
外部效度	新机构、时间、设备、人群是否验证？	开发集表现不可外推
人机协作	专业人员怎样理解、使用、推翻和接管？	形式“人在环”不够
临床后果	是否改变检查、治疗、延迟、并发症和患者结局？	中间指标不替终点
净伤害	假阳性、漏诊、过度诊断、警报疲劳、机会成本怎样？	收益与伤害必须同场
公平	谁受益、谁受损、亚组和可及是否变化？	平均值不结算少数
真实世界	规模部署后效果和罕见风险是否继续监测？	授权不是永久有效
版本	模型、数据、阈值、接口、工作流改变了吗？	版本变化可能改变证据对象

项目	必须回答	最低纪律
患者权利	能否理解 AI 参与、拒绝、申诉并继续获得服务？	交伦理法学与制度
公共卫生	监测提高是否真正改善干预、人群结局与公平？	更快信号≠更好健康
韧性	故障、网络攻击、供应商退出时能否继续医疗？	可用性属于临床安全的一部分

本表不是证据分数、临床准入算法或监管替代物。不同疾病、风险和技术需要不同研究设计。其作用只有一个：阻止证据从一个对象静默越级到另一个对象。

生成式人工智能使用声明

生成式人工智能参与了本稿的资料与文献线索整理、结构设计、部分文字起草与润色、交叉核查、引用线索核验和版式检查。AI 输出不作为经验数据、权威来源或独立证据；所有医学、公共卫生和健康系统事实均回到可核验的原始研究、正式指南、专业报告或监管材料。研究问题、理论命题、A／B／C／F 身份判断、规范承诺、材料取舍、引用采用与最终定稿均由作者审定，作者承担最终学术责任。生成式 AI 不列为作者。正式投稿时再依目标期刊要求补充具体工具、版本、日期与使用范围。

医学专门纪律：生成式 AI 生成的诊断、治疗、风险数字、药物建议、临床有效性或人群健康结论均不得因文本流畅而直接进入本文；任何定量医学主张必须回到相应研究对象、样本、设计、终点和限制。本文不提供个人医疗建议。

参考文献

[1] World Health Organization. Regulatory considerations on artificial intelligence for health. Geneva: WHO; 2023. ISBN 978-92-4-007887-1.

[2] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024. ISBN 978-92-4-008475-9.

[3] World Health Organization. Global patient safety action plan 2021–2030: towards eliminating avoidable harm in health care. Geneva: WHO; 2021. ISBN 978-92-4-003270-5.

[4] World Health Organization. Global patient safety report 2024. Geneva: WHO; 2024. ISBN 978-92-4-009545-8.

[5] U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices. Updated periodically; accessed 26 Aug 2026.

[6] U.S. Food and Drug Administration. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff. Final Guidance. 18 Dec 2025.

[7] U.S. Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff. Final Guidance. Issued 6 Jan 2026; re-issued 29 Jan 2026. January 2026 revisions included Level 2 minor corrections.

[8] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

- [9] Moons KGM, Damen JAA, 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. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
- [10] Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31:3283–3289. doi:10.1038/s41591-025-03953-8. Author Correction published 13 Jul 2026.
- [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. *Nat Med*. 2022;28:924–933. doi:10.1038/s41591-022-01772-9.
- [12] Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26:1364–1374. doi:10.1038/s41591-020-1034-x.
- [13] Cruz Rivera S, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26:1351–1363. doi:10.1038/s41591-020-1037-7.
- [14] Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554. doi:10.1136/bmj-2024-081554. Author-name correction: *BMJ*. 2025;388:r340. doi:10.1136/bmj.r340.
- [15] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. doi:10.1186/s12916-019-1466-7.
- [16] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574. doi:10.1177/0272989X06295361.
- [17] Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626.
- [18] Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969.
- [19] Hernström V, Josefsson V, Sartor H, et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. *Lancet Digit Health*. 2025;7(3):e175–e183. doi:10.1016/S2589-7500(24)00267-X.
- [20] Gommers J, Hernström V, Josefsson V, et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *Lancet*. 2026;407(10527):505–514. doi:10.1016/S0140-6736(25)02464-X.
- [21] Eisemann N, Bunk S, Mukama T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat Med*. 2025;31:917–924. doi:10.1038/s41591-024-03408-6.
- [22] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453. doi:10.1126/science.aax2342.
- [23] Pfohl SR, Cole-Lewis H, Sayres R, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat Med*. 2024;30:3590–3600. doi:10.1038/s41591-024-03258-2.
- [24] Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nat Med*. 2024;30:2838–2848. doi:10.1038/s41591-024-03113-4.
- [25] World Health Organization & World Bank. Tracking universal health coverage: 2025 global monitoring report. Geneva/Washington, DC; 2025.

- [26] World Health Organization. Health and care workforce – Global strategy on human resources for health: workforce 2030. Executive Board 156th session, B156/15. 2025.
- [27] World Health Organization. WHO upgrades its public health intelligence system to boost global health security: launch of EIOS 2.0. 13 Oct 2025.
- [28] World Health Organization. WHO Health Emergencies EPI-WIN webinar: artificial intelligence-supported listening to communities for cholera. 6 May 2026.
- [29] World Health Organization. A Global Health Strategy for 2025–2028: Fourteenth General Programme of Work. Geneva: WHO; 2025. ISBN 978-92-4-010101-2.
- [30] National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies (ECD7). Updated 9 Aug 2022.
- [31] Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med. 2021;4:54. doi:10.1038/s41746-021-00423-6.
- [32] National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Washington, DC: The National Academies Press; 2015. doi:10.17226/21794.
- [33] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56. doi:10.1038/s41591-018-0300-7.
- [34] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. doi:10.1186/s12916-019-1426-2.
- [35] Duggan MJ, Gervase J, Schoenbaum A, et al. Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency. JAMA Netw Open. 2025;8(2):e2460637. doi:10.1001/jamanetworkopen.2024.60637.
- [36] Olson KD, Meeker D, Troup M, et al. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. JAMA Netw Open. 2025;8(10):e2534976. doi:10.1001/jamanetworkopen.2025.34976.
- [37] Makar J, Abdelmalak J, Con D, Hafeez B, Garg M. Use of artificial intelligence improves colonoscopy performance in adenoma detection: a systematic review and meta-analysis. Gastrointest Endosc. 2025;101(1):68–81.e8. doi:10.1016/j.gie.2024.08.033.
- [38] World Health Organization. Epidemic Intelligence from Open Sources (EIOS) initiative. Current official initiative page; accessed 26 Aug 2026. By the end of 2025, 120 countries and 30 organizations and networks were leveraging the EIOS system.

版本说明

V1.1 在 V1.0 首个完整稿基础上完成专业审查与必要受控修订。本文的独立事实责任仍为临床有效性、临床净收益、患者安全、真实世界证据、公共卫生、人群健康与健康系统韧性；不替生命科学、AI／计算机科学、伦理法学、经济学、政治学或文明方向完成各自裁决。

本版本的主论证轴继续为：模型性能 → 外部验证 → 真实工作流与人机协作 → 临床行为 → 患者结局 → 真实世界持续监测 → 人群健康与健康系统后果。任何一步都必须说明证据对象、适用范围与停止点。本次受控修订主要包括：明确 Goh 等研究中模型单独表现为探索性次级分析；精确化 PRAIM 的前瞻性、多中心、真实世界、观察性非劣效实施设计及样本口径；区分 MASAI 随机入组与分析样本；校准 FDA 临床决策支持软件 2026 年最终指南的发布／再发布信息；登记 FUTURE-AI 作者名勘误；并移除外发稿中不必要的内部流程接力语言。主论证轴与 A／B／C／F 边界不变。V1.1 为当前专业研究基线。本次跨稿一致性同版回补仅同步文明方向稿版本、候选承载最小释义与外发导航表述，不改变医学专业结论。