卫生信息化国际发展动态

北京市卫生健康大数据与政策研究中心 2025.4.24

（三）人工智能评估

1.标题：缺血性中风管理的临床决策支持工具 GPT-4 的评估研究

来源： JMIR AI.

时间： 2025 年 3 月.

链接： https://doi.org/10.2196/60391.

概要：脑血管疾病是全球第二大最常见的死亡原因，也是造成残疾负担的主要原因之一。人工智能的进步有可能彻底改变医疗保健服务，尤其是在缺血性中风管理等关键决策场景中。本研究将通过对比急性缺血性中风管理的专家意见和真实世界结果，评估 GPT-4 在为急诊科神经科医生提供临床支持方面的有效性。研究回顾性地分析了 100 例急性卒中症状患者的队列，然后将每个病例独立提交给 GPT-4，GPT-4 给出关于治疗适当性、组织纤溶酶原激活剂的使用和血管内血栓切除术需求的比例建议。GPT-4 还评估了每位患者 90 天死亡概率，并阐明了其每项建议的原因，然后将这些建议与卒中专家的意见和实际治疗决策进行比较，其结果是：100 名患者队列中，GPT-4 的治疗建议与专家意见和真实世界治疗决策高度一致；GPT-4 在推荐血管内血栓切除术和组织纤溶酶原激活剂治疗也与实际决策一致。值得关注的是，在一定程度上，GPT-4 治疗建议比人类专家更积极，有 11 例 GPT-4 建议使用组织纤溶酶原激活剂，与专家意见相佐。对于死亡率预测，GPT-4 在其前 25 项高风险预测中准确识别了 13 例死亡中的 10 例（77%），优于 PRACTICE 和 PREMISE。本研究证明了 GPT-4 作为急性中风管理中可行的临床决策支持工具的潜力。它能够在不需要结构化数据输入的情况下提供可解释的建议，这与治疗医生的常规工作流程非常吻合。然而，更积极的治疗建议的趋势凸显了人类监督在临床决策中的重要性。未来的研究应侧重于前瞻性验证和探索将此类人工智能工具安全集成到临床实践中。

2.标题：评估人工智能在临床实践中的随机对照试验

来源：The Lancet Digital Health.

时间：2025年4月.

链 接 ： https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00047-5/fulltext.

概要：人工智能（AI）在过去5年的医疗应用中显著增长，在许多任务和专业中的表现已经与临床医生一样好，甚至更好，但是大多还是试验性的，缺乏在真实世界中的评估。这易导致AI模型中潜藏的偏差没被发现，在真实使用时给患者和临床医生造成严重风险。本研究将通过观察在临床实践中使用AI干预，进行机器学习的前瞻性随机对照试验评估，并对不同算法、不同群体和不同模式的差异和试验结果进行比较。研究重点分析了人工智能在改善护理管理、患者行为和症状以及临床决策效率方面的潜力，并确定了需要更多研究的领域。研究有助于利益相关者更好地理解AI的临床相关性和准备情况，并指导这个快速发展的领域的未来研究。研究表明，临床专业及相应机构对人工智能的兴趣越来越大，其中美国和中国在试验数量上处于领先地位，重点用于医学成像的深度学习，特别是在胃肠病学和放射学方面。研究突出了人工智能在加强护理管理、患者行为和症状以及临床决策方面的潜力，但是单中心试验的优势、少量人口统计学报告和不同运营效率的报告也引起人们对这些结果的普遍性和实用性的担忧，这种早期的成功可能受到发表偏倚的影响，其真正成功最终取决于它们对目标患者群体和环境的普遍性。为了更全面地了解AI在医疗保健中的真正影响和局限性，未来必须进行更多研究，包括关注多中心试验和纳入不同结果指标，尤其是与患者相关的结果。

（徐健编辑）

译文一：

缺血性中风管理的临床决策支持工具 GPT-4 的评估研究

Amit Haim Shmilovitch; Mark Katson; Michal Cohen-Shelly;
Shlomi Peretz; Dvir Aran; Shahar Shelly;徐健（译）

介绍

GPT-4 由 OpenAI 在 2023 年 3 月推出，标志着人工智能（AI）及其在包括医疗保健在内的各个领域的应用发展的一个重要里程碑。GPT-4 是 GPT 旗下的模型，体现了大语言模型（LLM）技术的进步。这项技术的基础架构涉及在广泛的数据集上进行训练，使模型能够作为“小样本学习器”运行。此功能使 GPT-4 能够适应新领域，并通过持续学习不断完善其性能。

在临床医学领域，GPT-4 等 LLM 的潜在应用特别有价值。这些模型有望作为医疗保健专业人员的支持工具，有助于有效地总结患者数据，协助决策过程，并可能提高医疗干预的准确性和速度。最近的研究强调了 GPT-4 在复杂医疗任务中的能力。值得注意的是，该模型已证明在类似于美国医师执照考试的考试中表现出熟练程度，达到或接近及格门槛的分数。此外，在以神经病学委员会考试题为模型建模的评估中，GPT-4 显示出很高的准确率，并且随着反复尝试而提高。

急性缺血性卒中（AIS）的管理在临床环境中是一项关键且对时间敏感的挑战。诊断和治疗 AIS 的方法需要综合信息，包括患者症状、体格和神经系统检查、病史和影像学结果。尽管美国心脏协会/美国卒中协会提供了既定的卒中管理指南，主治医师的判断仍然起着关键作用。临床表现的可变性和对决策的迫切需求凸显了 AI 辅助工具在这种情况下的潜在价值。此外，预测 AIS 的早期死亡率对于指导治疗决策、优化医疗保健环境中的资源分配、促进与患者及其家人的有效沟通、支持研究和临床试验以及促进质量改进计划至关重要。相应地，近年来已经为这项任务训练了几种传统的机器学习模型。

本研究将利用来自一家大型转诊医院急诊科（ED）的患者数据，重点关注出现中风症状的个体，以评估 GPT-4 在为 AIS 治疗提供准确临床决策方面的有效性。我们还评估了其预测 90 天死亡率结果的熟练程度。本研究的目的是量化像 GPT-4 这样的高级语言模型在多大程度上可以增强 AIS 管理中的临床决策过程。具体来说，我们假设 GPT-4 可以提供与人类专家相当的准确治疗建议和死亡率预测，从而可能有助于在急诊医学中最关键领域之一改善患者的预后。

方法

群组选择

这项回顾性研究包括来自兰巴姆保健校园急诊科的 100 例连续病例。2022 年 1 月至 2023 年 4 月期间接受治疗的所有患者均确诊为 AIS。纳入标准包括 18 岁以上的患者、美国国立卫生研究院卒中量表（NIHSS）评分为 5 分或更高（场外接受组织纤溶酶原激活剂[tPA]的患者 93 除外），并且从症状出现到接受脑部非对比计算机断层扫描（CT）不到 5 小时。所有纳入的患者在急诊室接受了脑部 CT、CT 血管造影和 CT 灌注平扫。该队列是专门选择的，因为它符合美国心脏协会急性卒中管理指南，使每位患者都是 tPA 和血管内血栓切除术（EVT）治疗的潜在候选人。共有 17 例不符合这些标准的患者被归类为“复杂”病例，其中临床情况需要额外考虑非指南治疗方案，并且需要评估个体患者的独特特征、病史和病情。对于每位患者，我们收集了从急诊科到达时的全面医疗记录，包括影像学检查结果，并将其从希伯来语翻译成英语。排除标准是临床数据不完整或卒中不是最终诊断的患者。

每位患者的临床数据包括人口统计学、病史、主诉、症状发作时间、体格和神经系统检查、NIHSS 评分、影像学结果（包括阿尔伯塔省卒中计划早期 CT 评分（如果可用）、接受的治疗和死亡率数据。一位经验丰富的卒中专家对结果不知情，审查了病例并在不治疗、tPA、EVT 或 tPA 和 EVT 的组合中做出了治疗决定。所有数据都经过去标识化处理，删除了标识符、名称和日期。

分析方法

该分析使用了 OpenAI 应用程序编程接口“创建聊天完成”方法，模型为 gpt-4-1106-preview。设置了默认参数（temperature=1;top_p=1;n=1），并使用 R（R 统计基金会）包装库 *open ai* 提交。完整的提示和示例在多媒体附录 1。为了评估 GPT-4 反应的可靠性，每个案例都进行了 5 次提交，以及一次没有伴随临床表现叙述的额外提交。对于每个治疗决定，GPT-4 都提供了叙述性解释。在 95%（475/500）的情况下，GPT-4 在请求的结构中返回响应，这些响应被 R 自动抓取。对于以范围形式提供的估计值，使用平均值。如果 GPT-4 提供了一个带有更大符号的数字（例如，>50），则该数字将记录为额外的 5。在 0.8%（4/500）的病例中，GPT-4 没有返回治疗决策的数字响应，在 8.6%（43/500）的响应中，它没有提供 90 天的死亡率估计值。

统计分析

GPT-4 的回答从 1 到 7 用于治疗决策，从 0 到 100 用于 90 天死亡率估计。计算 5 次重复的平均值。所有统计分析均使用 R（4.3.2 版）进行，使用基本 R 函数，预测受试者工作特征（ROC）1.18.5 和生存率 3.5.7。ROC 曲线被平滑。使用 psych2.3.12 库，使用线性加权 Cohen κ 系数测量治疗决策之间的一致性。

伦理考虑

本研究经兰巴姆保健校园赫尔辛基委员会（0156-24-D）批准作为回顾性分析。由于研究的回顾性和使用去标识化数据，放弃了知情同意的要求。所有患者信息在分析前都经过匿名化处理，并删除了所有标识符、姓名和日期，以确保隐私和机密性。由于这是一项使用现有临床数据的回顾性研究，因此没有向参与者提供补偿。该研究不涉及任何可能识别个体参与者的图像。这项研究是根据《赫尔辛基宣言》的原则进行的，并遵守所有相关的机构和国家研究伦理准则。

结果

患者人口统计学和临床数据

表 1.研究队列临床信息和人口统计数据

变量		简单案例（n=83）	复杂病例（n=17）
女性，n（%）		38(46)	7(41)
年龄（岁），中位数（IQR）		75.0(68.0-79.5)	71.0(65.0-77.0)
第一 NIHSSa，中位数（IQR）		12.0(8.5-16.5)	5.0(5.0-9.0)
到 CTb 的时间（小时），中位数（IQR）		1.8（I1.5-2.6）	4.45(3.0-5.2)
头颅 CT 结果，n（%）			
	左心室 c	48(58)	7(41)
	MCAd	47(57)	4(24)
	主成分分析 e	8(10)	4(24)
风险因素，n（%）			
	高血压	51(61)	10(59)
	DM 女	35(42)	3(18)

变量		简单案例（n=83）	复杂病例（n=17）
	血脂异常	36(43)	6(35)
	吸烟	11(13)	4(24)
	慢性肾病 g	11(13)	0(0)
	肥胖	5(6)	0(0)
	癌症	9(11)	1(6)
	HF 小时	7(8)	1(6)
	心律失常	19(23)	2(12)
	CADi 家族史	1(1)	0(0)
tPAj，n（%）		29(35)	7(41)
EVTk，n（%）		29(35)	1(6)
tPA+EVT，n（%）		12(14)	0(0)
90 天死亡率，n（%）		11(13)	2(12)
总死亡率，n（%）		17(20)	4(24)

[a]NIHSS：美国国立卫生研究院卒中量表。[b]CT：计算机断层扫描。[c]LVO：大血管闭塞。[d]MCA：大脑中动脉。[e]PCA：大脑后动脉。[f]DM：糖尿病。[g]CKD：慢性肾病。[h]HF：心力衰竭。[I]CAD：冠状动脉疾病。[j]tPA：组织纤溶酶原激活剂。[k]EVT：血管内血栓切除术。

我们从兰巴姆保健校园急诊室出现急性中风症状的连续 100 例患者中生成了一个队列。所有病例在急性卒中紧急情况下接受了全面的临床和放射学评估，并由神经科医生进行了全面评估（表 1 和图 1 对其中 78 名患者进行了血运重建治疗：36 名接受 tPA 治疗，30 名接受 EVT，12 名接受两者。在该队列中，13 名患者在 90 天内死亡，共 21 名。总体而言，17 例在不符合确切治疗指南时被归类为“复杂”，每个病例的数据包括人口统计学、NIHSS 评分、到达脑部 CT 的时间、症状发作以及文本脑影像结果的详细信息和风险因素，这些因素在入院 ED 时可作为病史获得（表 S1 在多媒体附录 2)。

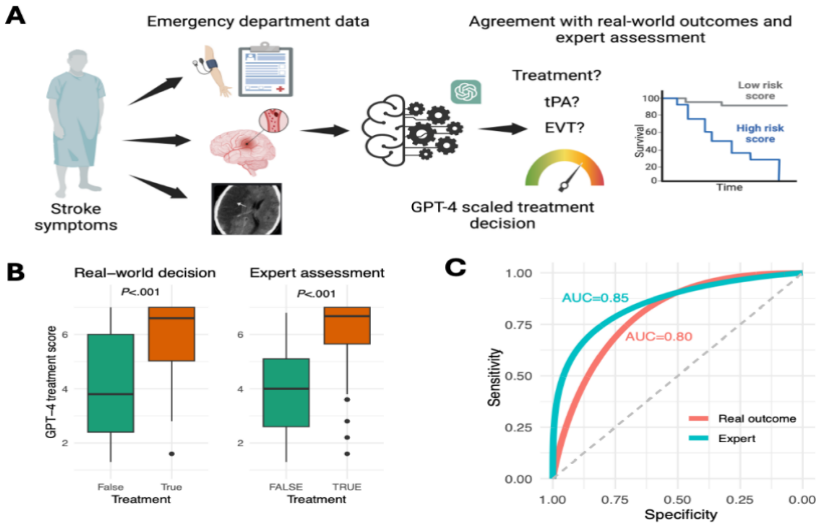

图 1. 研究设计和 GPT-4 性能评估

（A）研究设计图示，涉及 100 名连续的卒中患者，这些患者在到达急诊科时接受了全面的卒中检查，包括灌注、血管造影和脑部平扫 CT。记录临床资料、人口统计学、合并症和 CT 灌注结果。这些调查的文本报告被输入到 GPT-4API 中，该 API 被指示提供分数，表明是否治疗患者、是否使用 tPA、是否进行 EVT 以及 90 天死亡率的估计。

（B）箱形图显示用于治疗决策的 GPT-4 评估的平均分数（y 轴）。根据现实世界的决策和每个案例的专家评估进行比较（正确：治疗患者，错误：不治疗）。

（C）与真实世界决策和专家评估相比，GPT-4 治疗决策平均评分的 ROC 曲线和 AUC 评分。API：应用程序编程接口;AUC：曲线下面积;CT：计算机断层扫描;EVT：血管内血栓切除术;ROC：受试者工作特征;tPA：组织纤溶酶原激活剂。

一位对结局不知情的中风专家回顾性评价了每个病例。在其中 82 个案例中，专家的决定与实际实施的治疗一致。值得注意的是，专家建议不要治疗 11 名接受治疗的患者，并建议对 7 名未接受治疗的患者进行治疗。关于特异性治疗，在 61 例中观察到完全一致，尽管专家比在实践中观察到的更频繁地建议联合 tPA 和 EVT（Cohen κ =0.51，表示中等一致）。

GPT-4 临床决策

独立地，每个病例都使用 GPT-4 进行评估，生成从 1=不推荐干预到 7=强烈推荐（图 1 一个;表 S2 在多媒体附录 2）.为了解释 GPT-4 反应的可变性，每个案例都评估了 5 次。各次运行治疗评分的 Cohen κ 范围为 0.56 至 0.73。正如预期的那样，预定义的“复杂”案例在运行之间表现出显著更大的方差（*P*=.02）。

将 GPT-4 的治疗量表与专家决定和实际治疗进行比较，发现接受治疗的患者 GPT-4 的平均评分平均比未接受治疗的患者高 1.9 分（*P*<.001），与专家决定相比相差 2.1 分（*P*<.001;图 1 与实际决策相比，平均评分提供的 ROC 曲线下面积（AUC-ROC）为 0.80（95%CI0.69-0.91），与专家决策相比为 0.85（95%CI0.77-0.93）（图 1AUC 的这些平均分数高于每次独立运行（多媒体附录 3).此外，从 GPT-4 的分析中删除临床表现叙述导致 AUC 在实际决策中下降到 0.70，在专家决策中下降到 0.72（多媒体附录 3），强调了非结构化叙述数据在治疗决策中的重要性。同样，将 GPT-4 的温度设置为 0 导致现实世界和专家决策的 AUC 分别为 0.70 和 0.72，这表明需要让 GPT-4 有更多的创造力来获得更好的决策。

使用 4 分阈值，我们观察到 GPT-4 与现实世界治疗之间的 22 个分歧和 20 个与专家决定的分歧。值得注意的是，这些分歧中有很大一部分与专家和现实世界决策存在分歧的情况相吻合，在 30 个此类案例中，有 18 个（60%）显示了这种双重分歧。此外，复杂案例更容易出现差异，因为在 17 个复杂案例中，有 7 个与现实世界的决定不一致，5 个与专家决定的不同。专家检查了 GPT-4 生成的解释性文本，以了解模型与其盲法评估之间的所有差异，评估他们是否同意解释性文本作为原始模型输出的一部分是合乎逻辑的，可以被视为良好实践。在发生分歧的 20 个实例中，有 3 个案例，专家在仔细考虑了 GPT-4 的详细解释后，承认 GPT-4 的评估比他们最初的决定更可取。在另外 2 个案例中，专家承认 GPT-4 建议的方法确实是可以接受的，并且与可行的治疗方案一致。在专家不同意 GPT-4 推理的情况下，分歧主要围绕 3 个关键问题。首先，GPT-4 错误地将异常血管造影结果与临床表现联系起来。一个说明性案例是右侧大脑中动脉狭窄的患者，表现为右侧偏瘫（病例 94）。尽管这两个元素在解剖学上可能无关，但 GPT-4 错误地将它们联系起来。第二个值得注意的问题与伦理考虑有关，特别是在涉及活动性喉癌和认知能力下降患者的病例中。根据指南，患者被认为有资格接受治疗，但专家决定不继续治疗，因为预期寿命短且他正在接受姑息治疗（案例 14）。第三，与指南的偏差出现差异，尤其是在远端血栓切除术的情况下。例如，对于一名 96 岁的 M2 梗阻患者（被认为是远端血栓），GPT-4 建议不要治疗，这是既定指南;然而，专家呼吁继续进行血栓切除术，因为 NIHSS 评分高，并且根据个人经验，过去此类病例的结果很好（病例 54）。

在评估 GPT-4 选择最佳治疗方案的能力时，它在推荐 EVT 方面与现实世界的决策几乎完全一致：GPT-4 建议所有接受 EVT 治疗的患者（42/42，100%）使用 EVT（平均分>4）。专家建议对 55 名患者进行 EVT，其中 50 名患者也被 GPT-4 推荐使用 EVT，对应于实际决策的 AUC 为 0.94（95%CI0.89-0.98），专家的 AUC 为 0.95（95%CI：0.90-0.99）（图 2 对于 tPA 治疗，GPT-4 建

议接受 tPA 治疗的 38 名患者中有 79 名（48%）接受治疗，这表明与专家的共识更密切。在专家推荐的 41 名 tPA 患者中，GPT-4 同意 35 名（85%），对应于实际决策的 AUC 为 0.77（95%CI0.68-0.86）和专家的 AUC 为 0.82（95%CI0.73-0.90）（图 2B）。

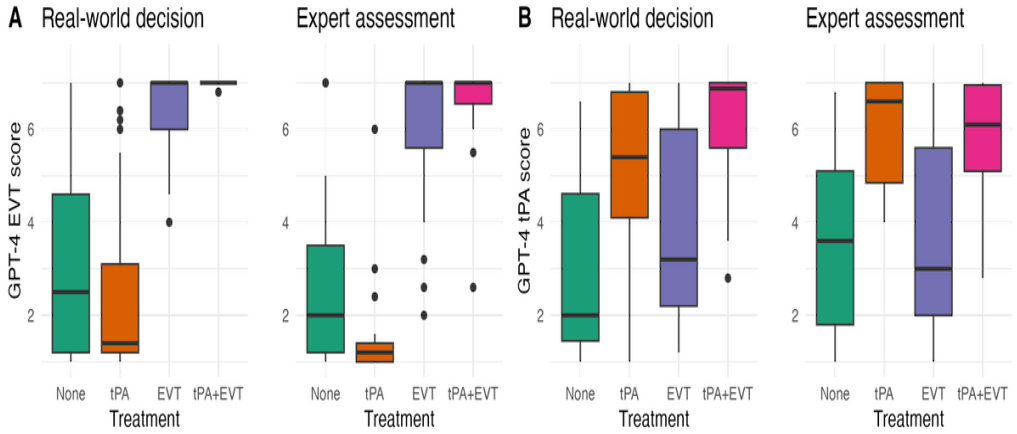

图 2. GPT-4 治疗类型评分

箱形图描述了 GPT-4 治疗类型评分，其中 y 轴表示概率评分（1-7 等级）。每个治疗类别都用颜色编码：绿色表示无干预，橙色表示 tPA，紫色表示 EVT，粉红色表示 tPA 和 EVT。（A）EVT 的 GPT-4 分数，按真实世界的决策和专家评估分层。（B）tPA 的 GPT-4 分数，按真实世界决策和专家评估分层。EVT：血管内血栓切除术;tPA：组织纤溶酶原激活剂。

死亡风险

我们进一步评估了 GPT-4 预测 90 天死亡率的能力。该模型估计 90 天内死亡的患者的平均死亡风险为 55.1%，而幸存者为 31.5%（P<.001），AUC 为 0.89（95%CI0.81-0.98;图 3A）。为了将这些结果置于上下文中，我们将 GPT-4 的性能与 2 个最近专门训练用于 90 天死亡率预测的机器学习模型的性能进行了比较。在我们的队列中，PRACTICE 模型的 AUC 为 0.70，明显低于 GPT-4 预测（对数秩 P 值=.02），而 PREMISE 模型的 AUC 达到 0.77（P=.07;图 3A）。这些比较强调了 GPT-4 在死亡率风险评估方面的显着准确性，优于经过训练的专业预测模型。

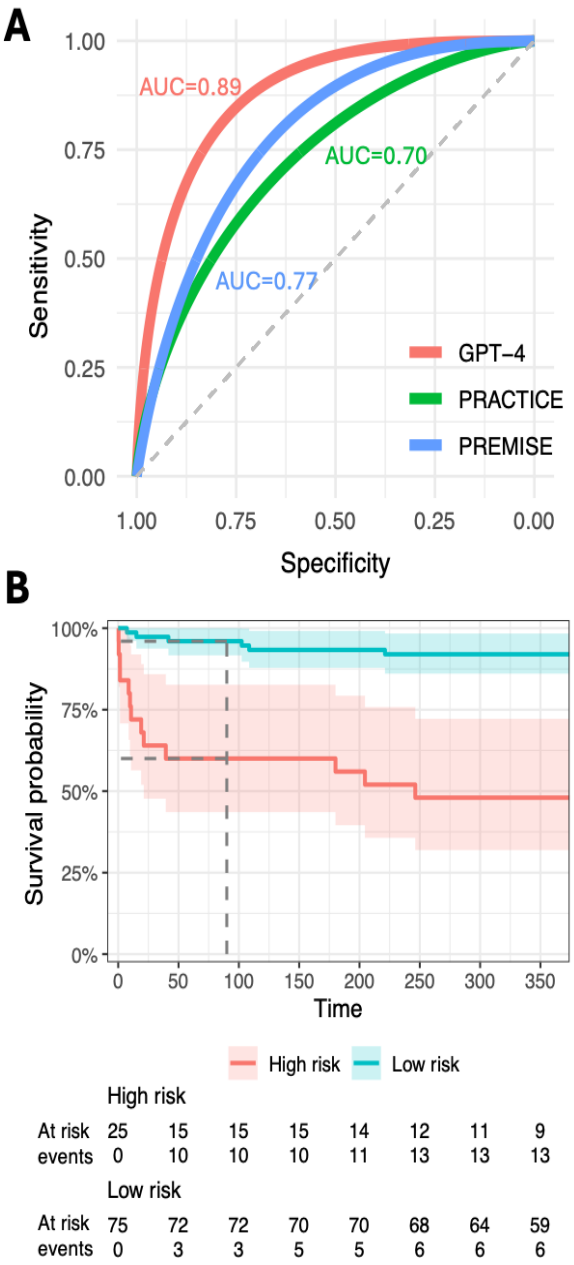

图 3. GPT-4 死亡率预测

（A）GPT-4（红色）、PRACTICE（绿色）和 PREMISE（蓝色）估计 90 天死亡率的 ROC 曲线。（B）Kaplan-Meier 图，根据 GPT-4 的 90 天死亡率估计将个体分为低风险和高风险类别。AUC：曲线下面积;ROC：受试者工作特征。

为了识别高危患者，我们在队列的前 25%处设定了一个阈值，这对应于 41%的预测死亡风险临界值。在这个高危人群中，有 10 名患者在入院后 90 天内去世，另有 3 名患者在随后的一年内去世（图 3B）。相反，在其余 75 名被归类为低风险的患者中，只有 3 例死亡发生在 90 天内，第一年总共发生 6 例。计算出的风险比为 6.98（95%CI2.88-16.9;P<.001），增强了该模型根据

患者的死亡风险对患者进行有效分层的能力。

讨论

研究展示了 GPT-4 作为 AIS 管理中临床决策支持工具的潜力。我们的主要发现表明，GPT-4 的治疗建议与专家意见（AUC0.85）和现实世界的决策（AUC0.80）密切相关。值得注意的是，GPT-4 在预测 90 天死亡率方面表现出很高的准确性（AUC0.89），优于专门的机器学习模型。

AIS 是全球死亡和残疾的主要原因。在无法获得专业卒中病房或合格医生的地区，卒中护理的紧迫性尤为重要。GPT-4 能够在现有治疗程序中无缝运行，仅依赖于常规图表信息，这使得它对于资源贫乏地区的快速分类很有价值。这种可访问性可以使高级别医疗咨询民主化，将专家级决策扩展到资源不足的医疗保健机构。

在我们的研究中，GPT-4 在预测接受血管内治疗的 AIS 患者的 90 天死亡率方面表现出很高的准确性。该模型使用了多种临床和影像学变量，与休斯顿动脉内治疗、休斯顿动脉内治疗 2、PREMISE 和 PRACTICE 等现有模型相比，提供了更全面的方法。与依赖结构化数据的传统医疗保健预测模型不同，GPT-4 提供了基于叙述文本的建议。我们的分析强调了非结构化数据的重要性，当排除叙述性临床表现时，预测准确性的下降证明了这一点。这展示了 GPT-4 以符合临床信息自然流的方式处理复杂医疗数据的能力。

在医疗保健领域部署 GPT-4 等 AI 模型的一个关键方面是其决策过程的透明度和可解释性。虽然 GPT-4 的自然语言输出可以给人一种可解释性的印象，但这些可能不一定反映真正可靠的推理过程。我们的分析集中在 GPT-4 基本原理的表面价值上，专家评审员认为这些理由很有见地。但是，我们承认可能会有令人信服但有缺陷的解释，这是 LLM 的一个已知局限性。这凸显了对此类模型输出进行批判性评估和谨慎解释的重要性，尤其是在高风险的医疗决策环境中。在将 AI 系统更广泛地整合到临床实践中之前，需要进行持续的研究来解决其推理过程的透明度和可靠性问题。

尽管结果令人鼓舞，但我们的研究也有一些局限性。我们必须承认应用 GPT-4 时存在一些挑战，尤其是在它评估伦理问题的能力方面。该模型在解决医疗决策固有的细微而复杂的伦理考虑方面可能面临困难。这种限制强调了在敏感的医疗保健环境中部署 GPT-4 等 AI 工具时需要谨慎和补充的人工监督。“幻觉”或错误输出的发生是另一个问题，尽管我们证明运行多次评估可以减轻这种风险。未来的研究应侧重于改进这些方法以进一步减少不准确。

另一个考虑因素是这些发现的普遍性。虽然这些建议可能部分反映了临床医生在临床记录

中编码的直觉，但我们的分析表明，该模型的评估超出了单纯的解释。在 GPT-4 建议与实际治疗决策和专家评估之间观察到的差异表明，该模型能够根据提供的数据进行独立评估。此外，临床表现笔记和影像学报告解释（表 S1 多媒体附录 2）没有明确传达临床医生的治疗偏好或直觉，这表明 GPT-4 不仅仅是反刍临床医生的思维过程。另一个可能的局限性是该研究的排除标准，特别是回顾性排除了临床数据不完整的患者或最终被诊断出患有中风以外的疾病的患者。虽然这些排除对于确保研究侧重于准确诊断的 AIS 病例是必要的，其中 GPT-4 决策支持能力可能最相关，但我们承认这种方法可能会限制我们的研究结果在更广泛的临床环境中的普遍性。在现实世界中，临床医生在做出治疗决策时经常面临诊断的不确定性和不完整的信息。最后，我们的研究是在具有特定患者群体的单个中心进行的。有必要在不同环境和更大人群中进行进一步研究，以验证 GPT-4 在各种临床环境中的有效性和适用性。

总之，我们的研究介绍了一种使用 GPT-4 进行中风管理临床决策支持的开创性方法。该模型已显示出处理叙述性文本、提供可解释的建议和增强医疗决策的潜力。随着我们继续探索和完善这项技术，它有望改变患者护理并改善最关键医学领域之一的结果。

译文二：

评估人工智能在临床实践中的随机对照试验

Ryan Han, Julián N Acosta, Zahra Shakeri, John P A Ioannidis,
Eric J Topol, Pranav Rajpurkar, 徐健（译）

介绍

人工智能（AI）在医疗保健中的使用在过去 5 年中取得了显著增长，一些出版物报告称，医疗 AI 模型在许多任务和专业中的表现与临床医生一样好或更好;然而，其中许多模型仅在真实世界临床环境之外使用替代品进行了回顾性测试。在美国食品和药物管理局批准或批准的近 300 种人工智能医疗设备中，只有少数经过了前瞻性随机对照试验（RCT）的评估。

AI 系统真实世界评估的稀缺性导致了巨大的不确定性，包括对患者和临床医生造成重大风险的可能性。这种风险的一个例子是广泛使用的脓毒症模型，该模型被发现的性能比其开发人员报告的性能“差得多”，由于警报不正确或不相关，导致“警报疲劳负担沉重”。人工智能在前瞻性部署时表现更差的情况可能并不少见，而在临床环境中采用人工智能系统的困难可能会进一步阻碍重要结果方面的任何潜在好处。此外，如果没有真实世界的评估，AI 模型的偏差可能仍然无法被发现，这可能会无意中导致健康结果的差异。

为了更清楚地了解医疗保健领域的 AI 前景，本范围综述旨在检查临床实践中使用的 AI 算法的 RCT 状态。尽管有几篇系统综述已经就这一主题进行了研究，我们的范围综述更新了截至 2023 年底发表的许多新试验的证据，因为自 2021 年以来发表的试验数量增加了一倍多。我们的范围审查还引入了新的纳入标准。具体来说，我们要求 AI 干预反映机器学习的当前进步，并集成到临床团队完成的实际患者管理中。这种对具有临床意义的 AI 应用的严格关注确保了我们的审查与为医疗实践提供信息密切相关。此外，我们的综述独特地检查了详细的分析，突出了算法的多样性、不同群体的比较、模式的差异以及试验结果的性质。这种区别使本范围综述与早期的系统综述不同，后者主要集中在评估总体证据、方法学的质量或统计严谨性。我们的分析考察了人工智能在改善护理管理、患者行为和症状以及临床决策效率方面的潜力，并确定了需要更多研究的领域。我们的目标是帮助利益相关者更好地理解 AI 的临床相关性和准备情况，并指导这个快速发展的领域的未来研究。

方法

检索策略和选择标准

我们系统检索了 PubMed、SCOPUS、CENTRAL 和国际临床试验注册平台（International Clinical Trials Registry Platform），以查找 2018 年 1 月 1 日至 2023 年 11 月 14 日期间发表的相关研究。选择这个时间表是为了与现代 AI 模型开始在试验中发挥重要作用的时代相吻合。我们使用了自由文本检索词，例如“artificial intelligence”、“clinician”和“clinical trial”。详细的检索策略可在附录（第 3-7 页）中找到。此外，我们手动审查了相关出版物的参考文献，以查找更多文章。

我们的纳入标准特定于满足以下条件的随机对照试验：干预纳入了大量人工智能组件，我们将其定义为非线性计算模型（即机器学习组件，包括但不限于决策树、神经网络等）;干预被整合到临床实践中，从而影响临床团队对患者的健康管理;结果作为全文文章发表在同行评审的英文期刊上。我们排除了评估线性风险评分的研究，例如 logistic 回归、二级研究、摘要和未纳入临床实践的干预措施。本范围综述遵循 PRISMA 范围界定审查指南（附录第 8-9 页），并且该范围审查的方案已在 PROSPERO（CRD42022326955）注册。

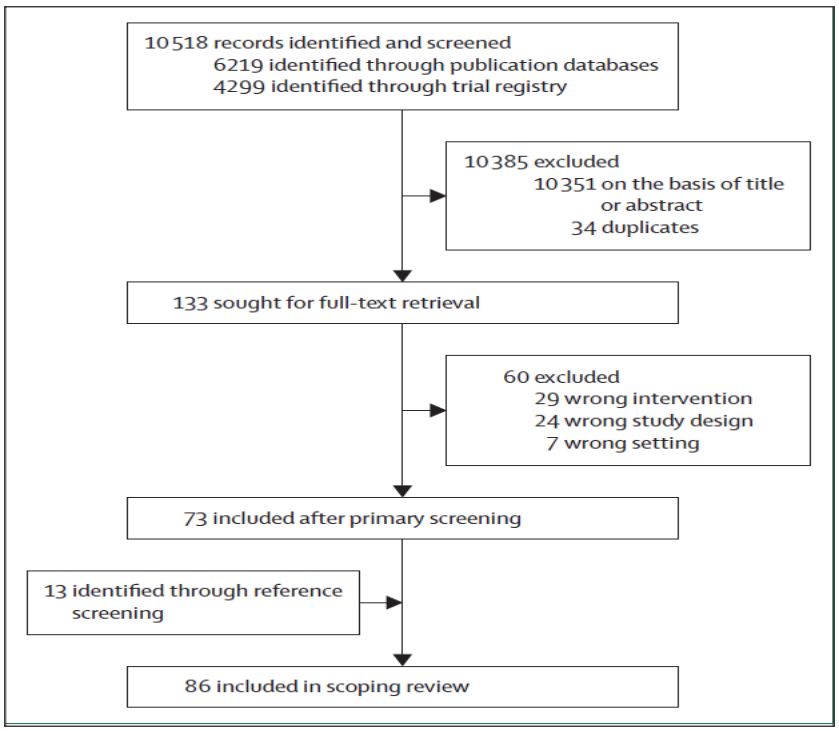

图 1 研究文献选择

数据分析

为了确保搜索结果的质量，我们使用 CovidenceReview 软件来筛选出版物标题和摘要。两名独立研究者（RH 和 JNA）进行了初步筛选，然后对筛选的论文进行了全文审查。符合条件的论文由一名研究者在 Google 表格中完成数据提取，然后由第二名研究者（RH 或 JNA）进行验证。通过与第三位评价员（PR）讨论解决任何差异。

我们提取了研究层面的信息，包括研究地点、受试者特征、临床任务、主要结果、时间效率结果、对照和结果，以及所用 AI 的类型和来源。此外，我们按主要结果组（诊断率或性能、临床决策、患者行为和症状以及护理管理）、临床领域或专业以及 AI 使用的数据模式对研究进行分类。

我们没有尝试联系研究作者以获取更多或不确定的信息。由于任务和结果的预期异质性，我们没有进行正式的 meta 分析。相反，我们提供了简单的描述性统计数据，以提供合格试验特征的概述。

结果

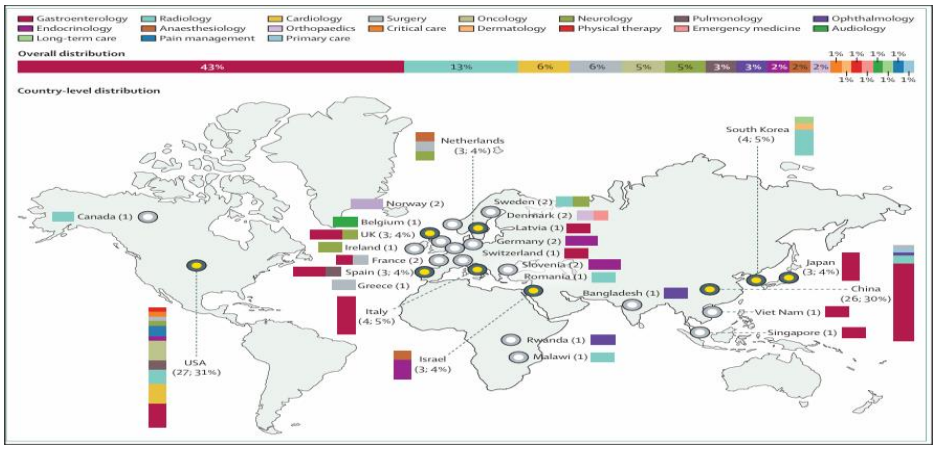

图 2：人工智能在不同国家和专业的临床实践中的随机对照试验

挪威、法国、瑞典、丹麦、德国和斯洛文尼亚各占分布的 2%。加拿大、比利时、爱尔兰、希腊、拉脱维亚、瑞士、罗马尼亚、孟加拉国、卢旺达、马拉维、越南和新加坡各占 1%。

我们的电子检索检索到 6219 条研究记录和 4299 份试验注册，去重后共得到 10484 条记录（图 1）。经过标题和摘要筛选，保留 133 篇文章进行全文审查。其中，60 例被排除在外，初步筛选后剩下 73 项研究。通过二次参考文献筛选确定了另外 13 篇文章，因此我们的范围综述中共纳入了 86 篇独特的 RCT。附录（p2）中提供了所有纳入研究的参考文献和特征。

在 86 项 RCT 中，37 项（43%）与胃肠病学有关，11 项（13%）与放射学有关，5 项（6%）与外科有关，5 项（6%）与心脏病学有关。胃肠病学的试验以其一致性而著称，所有试验都在支持临床医生的辅助设置中测试基于视频的深度学习算法，除一项试验外，所有试验都测量了与诊断率或性能相关的主要结果（检出率、漏诊率等）。37 项胃肠病学的试验中有 24 项（65%）仅由四组进行(8项试验来自武汉大学,6项来自WisionAI,6项来自Medtronic,4项来自Fujifilm)。

86 项 RCT 中有 79 项(92%)是在同一个国家进行的,其中美国进行的试验最多(27 项[31%])，其次是中国（26 项[30%]）。在美国进行的试验分布在各个专业，而在中国进行的 26 项试验中有 21 项（81%）主要与胃肠病学有关。在多个国家进行的试验主要涉及欧洲国家（7 项中的 6 项[86%]）。图 2 突出显示了试验在不同国家和专业之间的分布。

试验主要在单个中心进行(86 例中的 54 例[63%])，最终分析中位数为 359 例患者(IQR150-1050）。在 86 项试验中，83 项（97%）报告了受试者的平均或中位年龄，中位年龄为 57·3 岁（范围 0·0034-78;IQR49·9－62·0）。同样，在 86 项试验中，有 83 项（97%）报告了性别，其中男性参与者的中位数为 48·9%（范围 0-89·2;IQR45·4－54·2）。22 项试验报告了种族或族裔，其中 18 项（82%）来自美国。在这些试验中，白人（非西班牙裔或拉丁裔）参与者的中位百分比为 70·5%（范围 0-98·4;IQR35·0－81·8）。只有 3 项在中国和 1 项在韩国的试验明确地报告了一个种族：分别是中国汉人和亚洲人。

在 2021 年初以来发表的 63 项试验中，12 项（19%）引用了 2020 年 CONSORT-AI 报告指南，用于评估 AI 干预的临床试验。[16]

大约一半（86 项中的 46 项[54%]）试验的主要结果与诊断率或执行率有关，例如检出率或平均绝对误差。其他主要结果根据护理管理（18[21%]）、患者行为和症状（15[17%]）和临床决策（7[8%]）进行分组。表 1 总结了结果和结果类型的分布。

表 1 人工智能临床实践随机对照试验的主要结果和类型

评估内容	统计学上显著的改善	无统计学意义效应	表现出非劣效性	具有统计学意义的恶化	总计
护理管理	15	1	2	.	18
临床决策	6	1	.	.	7
诊断率或性能	34	10	1	1	46
患者行为和症状	10	3	2	.	15
总计	65	15	5	1	86

数据为 n。

18 项 RCT 评估了 AI 干预对护理管理质量指标的影响，为人工智能在临床实践中的使用提供了以结果为导向的视图。例如，用于胰岛素剂量和低血压监测的 AI 系统已被证明可以分别改善

患者在血糖和血压目标范围内的平均时间。同样，评估 AI 系统用于放射治疗和前列腺近距离放射治疗的试验也通过其降低急症护理率和前列腺肿瘤体积的能力进行评估。

15 个 AI 系统还评估了它们对患者行为和症状的影响。例如，一项试验报告称，与让患者等待临床医生进行分级相比，立即向患者提供 AI 生成的糖尿病视网膜病变风险预测可以提高转诊依从性。另一项试验报告称，与无辅助临床医生相比，采用伤害感受监测系统能够降低患者的术后疼痛评分。这些试验强调了 AI 干预对患者体验产生直接影响的潜力。

七项试验还测量了 AI 系统影响临床决策的能力。例如，据报道，癌症患者 AI 死亡率预测的可用性增加了肿瘤学家和患者之间严重疾病对话的数量。相比之下，采用 AI 系统来识别中风高危心房颤动患者并没有增加新的抗凝剂处方。[26] 这些研究探索了 AI 预测为临床医生的判断提供协作信息的潜力。

86 项试验中有 59 项（69%）评估了用于医学成像的深度学习系统。值得注意的是，所评估的医学成像系统主要是基于视频的（59 个中的 42 个[71%]），而不是基于图像的（59 个中的 17 个[29%]）。这种效果主要是由大量的内窥镜试验（42 项中的 34 项[81%]）驱动的。在成像之外，AI 系统还对结构化数据进行作，例如来自电子健康记录（27 个中的 14 个[52%]）、波形数据（27 个中的 10 个[37%]）和自由文本（27 个中的 3 个[11%]）。这些系统混合使用决策树（27 个中的 6 个[22%]）、神经网络（27 个中的 2 个[7%]）、强化学习（27 个中的 2 个[7%]）、基于大小写的推理（27 个中的 2 个[7%]）、贝叶斯分类器（27 个中的 1 个[4%]）和未指定的机器学习（27 个中的 14 个[52%]）。

大多数在医学成像上运行的系统（59 个[85%]中的 50 个[85%]）是在临床医生的辅助设置中进行评估的，而基于结构化数据的模型往往与常规护理（14 个中的 12 个[86%]进行比较）。模型主要在工业界开发（86 个中的 47 个[55%]），其次是学术界（86 个中的 35 个[41%]），其余四个模型具有混合或未说明的来源。

表 2 临床实践中 AI 随机对照试验的主要结果结果和组比较

评估内容	统计学上显著的改善	无统计学意义效应	表现出非劣效性	具有统计学意义的恶化	总计
AI 与临床医生	3	1	3	1	8
AI 与常规护理	16	4	.	.	20
AI 辅助临床医生与独立临床医生	46	10	2	.	58
总计	65	15	5	1	86

数据：n；AI：人工智能

表 2 总结了结果和组比较的分布。在 86 项试验中，81 项试图显示改善，5 项使用非劣效性

设计。在旨在显示改善的 81 项试验中，有 65 项（80%）报告了其主要结果的显著改善。其中 46 项（71%）指出，与无辅助临床医生相比，AI 辅助临床医生有所改善，16 项（25%）指出与常规护理相比，AI 系统有所改善，3 项（5%）报告独立 AI 系统的性能优于临床医生。

在采用非劣效性设计的 5 项试验中，3 项确定了独立 AI 系统和临床医生之间的非劣效性，2 项确定了辅助和独立临床医生之间的非劣效性。因此，86 项试验中有 70 项（81%）报告了其主要结果的良好结果。在胃肠病学亚群中观察到类似的成功率，37 项试验中有 28 项（76%）报告了显著改善，一项（3%）显示非劣效性，总体成功率为 78.4%。

16 项主要结果结果为阴性的 RCT 包括 10 项试验，这些试验没有显示与无辅助临床医生相比，辅助临床医生有所改善，4 项试验没有显示 AI 系统与常规护理相比有所改善，1 项试验没有显示与临床医生相比，独立 AI 系统有所改善。一项试验还报告说，独立的 AI 系统的性能明显差于临床医生;然而，这 16 项试验中有 8 项（50%）报告了次要结果的显著改善。

86 项试验中有 52 项（60%）还报告了手术时间测量，结果各不相同。大约三分之一的试验（52 项中的 18 项[35%]）报告手术时间显著减少（$p<0.05$）;然而，大约四分之一（52 人中的 13[25%]）报告手术时间显著增加（$P<0.05$）。其余 21 项试验中的 40 项（52%）发现手术时间测量没有显著变化。

胃肠病学是这些结果的主要贡献者，有 32 项试验涉及手术时间测量。这些结果各不相同，2 项试验（6%）指出手术时间减少，12 项试验（38%）报告手术时间增加，其余 18 项试验（56%）观察到没有显著影响。所有 5 项放射学试验和所有 3 项眼科试验都报告了手术时间的显著缩短。在其他专业中，通常考虑手术时间方面，两个或更少的试验。

讨论

这份对 AIRCT 出版物的范围审查揭示了几个值得注意的趋势和对临床实践中 AI 系统开发和实施的影响。试验在临床专业和地点的分布突出了 AIRCT 集中在胃肠病学、放射学、外科和心脏病学。值得注意的是，与专科护理相比，对初级保健的关注较少，这表明未来研究的潜在领域。试验的地理分布揭示了单一国家研究的优势，大多数试验来自美国，其次是中国。2023 年对人工智能和机器学习支持的设备试验注册的系统评价发现，专业和地域分布相似，还指出了国家试验的优势。然而，该范围综述还发现了不同的趋势，中国在试验注册方面处于领先地位，放射学是最常见的专业。这一发现表明需要更多的国际合作和多中心试验，以确保 AI 系统在不同人群和医疗保健系统中的普遍性。

单中心试验占主导地位，中位数为 359 名患者，这表明 AI 医疗保健试验通常选择较小的受控环境;然而，很少有人口统计报告，特别是关于种族和民族的报告，引起了人们对这些研究的代表性的担忧。CONSORT-AI 报告指南的引用频率低，进一步强调了试验方法需要更高的透明度。这种透明度将增强对试验对更广泛人群适用性的理解，因为纳入标准、设置和随访持续时间等因素会严重影响结果的普遍性。未来的试验应优先考虑全面的报告和受试者的多样性，以增强其结果的外部有效性。

在随机对照试验中评估的 AI 应用中，使用深度学习系统进行医学成像，尤其是在基于视频的系统中，是一个普遍的趋势。这一趋势在评估基于视频的胃肠病学干预的大量试验中很明显，这与基于图像的放射学算法在学术文献和监管许可中的主导地位形成鲜明对比。对于基于图像的放射学算法，除 RCT 之外的其他设计可能最适合解决诊断准确性问题。配对设计研究允许比较同一个体的诊断性能，消除所有混杂因素;然而，在胃肠病学应用中，例如腺瘤检测，配对设计是不可行的，因为检测到的病变通常会被去除。这一趋势似乎是由占大多数基于视频的胃肠病学试验的少数群体推动的，这表明临床 AI 试验领域在研究人员、试验设计和结果测量方面仍然是同质的。然而，使用结构化数据（如电子健康记录和波形数据）的系统混合使用了决策树、神经网络、强化学习和其他机器学习技术。这种多样化的模型和数据源显示了 AI 应对不同医疗保健挑战的适应性。需要更多的研究来评估将临床背景（多种模式）或临床先验（多个时间点）纳入决策的 AI 系统的效果，因为这些因素对许多临床任务至关重要。

我们的成功率与用于医疗干预和医疗保健 AI 系统的 RCT 历史综述的成功率之间的差异可归因于我们对 AI 和临床实践的具体定义，其中排除了没有临床整合和非线性 AI 的研究，以及我们更新的检索策略，其中包括一些新的和以前被忽视的试验。我们的综述将考虑窗口延长至 2023 年，因此与之前的综述相比，捕捉了这一快速发展领域一年多的进步和大量近期试验。尽管取得了这些有利的结果，但 AI 应用的普遍性仍不确定。指定 AI 训练数据是来自相同还是不同的机构，这对于试验至关重要。此外，比较在内部和外部测试环境中进行的 RCT 的分析可以为 AI 性能的普遍性提供有价值的见解。此外，应根据该领域的起步阶段和发表偏倚的可能性来看待对这一成功率的解释。2023 年的一项系统评价确定了 627 项在 ClinicalTrials.gov 上注册的人工智能技术试验，但只有 9 项（1%）被轻松确定为已发表。在被列为正在进行或没有发布结果的试验中，阴性结果的数量是未知的，这会导致其完成或结果的发布和发布延迟。因此，发表偏倚对临床实践中 AI 的整体效果和有效性的有效解释构成了重大威胁。

大多数试验评估了与诊断率或性能相关的结果干预措施。尽管此类试验为临床 AI 系统的前

瞻性技术性能提供了令人信服的证据，但这些证据可能无法准确反映 AI 系统对患者护理的整体影响，因为高灵敏度和特异性不一定会转化为改善患者预后。例如，2023 年对 21 项结肠镜检查试验的系统评价发现，尽管 AI 辅助有助于提高息肉检测率，但它并未显著提高临床关键晚期腺瘤的检测率。更一般地说，诊断性能和其他 AI 试验的统计学上有利的结果不一定会转化为具有临床意义的益处。一些试验评估了人工智能系统对护理管理质量指标、患者行为和症状以及临床决策的影响。这些不同的结果衡量标准反映了 AI 系统影响临床实践的各种方式，从提高护理质量到增强患者体验和为临床判断提供信息。为了更好地评估 AI 算法在医疗保健中的真正价值，真实世界的证据必须关注具有临床意义的结果，例如症状和治疗需求，以及生存等长期结果。此外，更大规模的证据将允许更好地评估这些结局的益处的绝对大小是否是实质性的。

在运营效率方面，结果因专业而异，大量试验报告运营时间增加或减少。这一发现突出了 AI 系统根据特定应用和环境简化临床工作流程或使其复杂化的潜力。鉴于这种复杂性，AI 工具的成功采用将取决于运营效率、成本效益和所需的培训水平等因素，以及性能。因此，未来的研究不仅应关注临床结果，还应关注实施的这些多方面，以便更全面地了解 AI 对医疗保健服务的影响。

总之，临床实践中关于 AI 的 RCT 的现有情况表明，人们对在一系列临床专业和机构应用 AI 的兴趣越来越大。大多数试验报告了积极的结果，突出了人工智能在加强护理管理、患者行为和症状以及临床决策方面的潜力，但这种早期的成功应该受到发表偏倚的可能性的影响。AI 应用程序的真正成功最终取决于它们对目标患者群体和环境的普遍性，而 Standing Together 倡议等工作在这一主题上提供了宝贵的指导。为了更全面地了解 AI 在医疗保健中的真正影响和局限性，必须进行更多研究，包括关注多中心试验和纳入不同的结果指标，尤其是与患者相关的结果。

此范围审查有两个重要的限制。首先，仅以英文检索相关研究。这种语言限制可能排除了以其他语言发表的相关试验，这可能会限制我们研究结果的全面性和普遍性。其次，尽管将考虑窗口延长至 2023 年，但我们的综述并未涉及试验偏倚风险的最新趋势。鉴于随机对照试验的不断涌入，未来的系统评价应解决试验偏倚风险的趋势（例如，使用 Cochrane 偏倚风险和其他相关工具），并提供更深入的报告透明度分析（CONSORT-AI）。