·解　读·

基于人工智能的临床辅助决策系统早期临床评价研究的报告规范（DECIDE-AI）解读

陈泞夙[1,2]，赵凯[2]，薛心雨[2]，齐亚娜[2]，喻佳洁[1,2]

1. 四川大学华西医院临床营养科（成都 610041）
2. 四川大学华西医院临床流行病学与循证医学研究中心/中国循证医学中心（成都 610041）

【摘要】 近年来人工智能在医疗服务领域应用愈发广泛，基于人工智能的临床决策辅助系统是其主要应用形式之一。基于人工智能临床决策辅助系统的早期临床评价介于临床前开发（计算机模拟）、离线验证和临床试验之间，但目前少有人工智能相关临床研究涉及人因学评价，且缺少与人工智能系统运行环境、用户特征、选择过程及算法识别等方面的报告。为缩小人工智能辅助决策系统在开发与实际临床应用间的差距，提高人工智能系统临床研究的透明性和规范性，2022 年，*BMJ* 发表了基于人工智能的临床辅助决策系统早期临床评价研究的报告规范（DECIDE-AI）。本文就指南的制订背景、制订过程和重点内容进行解读，以期促进该报告规范在国内研究人员中的理解与应用。

【关键词】 报告规范；人工智能；辅助决策系统；早期临床评价；解读

Interpretation of the DECIDE-AI guideline: a reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence

CHEN Ningsu[1,2], ZHAO Kai[2], XUE Xinyu[2], QI Yana[2], YU Jiajie[1,2]

1. Department of Clinical Nutrition, West China Hospital of Sichuan University, Chengdu 610041, P. R. China
2. Research Center of Clinical Epidemiology and Evidence Based Medicine/Chinese Cochrane Centre, West China Hospital of Sichuan University, Chengdu 610041, P. R. China
Corresponding author: YU Jiajie, Email: 2003xiong@163.com

【**Abstract**】 Artificial intelligence has been extensively applied in healthcare services recently, and clinical decision support systems driven by artificial intelligence are one of the applications. Early-stage clinical evaluation of artificial intelligence (AI)-based clinical decision support systems lies between preclinical development (in silico), offline validation, and large-scale trials, but few AI-related clinical studies have addressed human factors evaluations and reported the implementation environment, user characteristics, selection process and algorithm identification of AI systems. In order to bridge the development-to-implementation gap in clinical artificial intelligence and to promote the transparent and standardized reporting of early-stage clinical studies of AI-based decision support systems. A reporting guideline for the developmental and exploratory clinical investigations of decision support systems driven by artificial intelligence (DECIDE-AI) was published in 2022. This paper aimed to interpret the background, development process and key items of the DECIDE-AI guideline and promote its understanding as well as dissemination in China.

【**Key words**】 Reporting guideline; Artificial intelligence; Decision support systems; Early-stage clinical evaluation; Interpretation

1　DECIDE-AI 的制订背景

近年来，人工智能（artificial intelligence，AI）在医疗服务领域的应用愈发广泛，基于 AI 的临床决策辅助系统（以下简称“AI 系统”）是其主要应用形式之一。研究表明：AI 系统在临床前开发阶段或计算机生物模拟阶段已展现出可比拟人类专家的良好性能[1]，但少有证据证明其在临床实际应用中能改善医生活动和患者结局[2,3]。目前，AI 系统的开发与应用间尚存在的“AI 鸿沟”[4]，即主要强调人工智能算法的数学性能，忽略了人工智能系统、

DOI：10.7507/1672-2531.202401188
基金项目：四川省自然科学基金项目（编号：2023NSFC0520）
通信作者：喻佳洁，Email：2003xiong@163.com

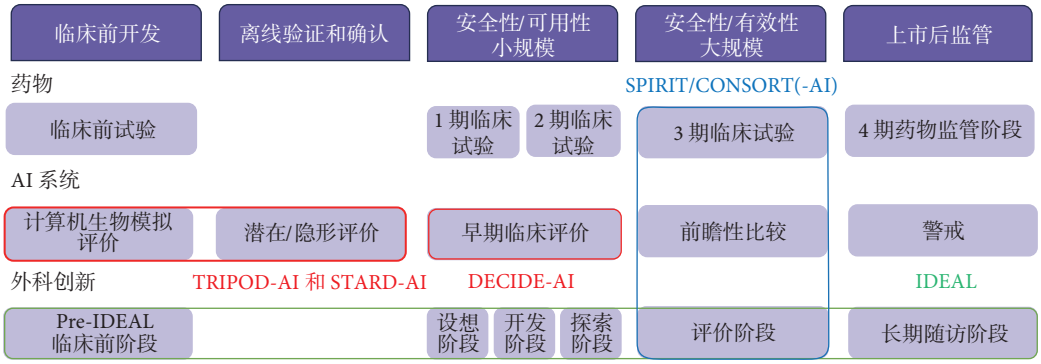

图 1　药物、医疗卫生领域中人工智能和外科创新的发展路径比较

彩色线条代表报告指南，其中部分是针对特定研究设计的（TRIPOD-AI、STARD-AI、SPIRIT/CONSORT、SPIRIT/CONSORT-AI），部分是针对特定阶段的（DECIDE-AI、IDEAL）。

用户和实施环境间相互作用对其实际应用的影响。将人工智能系统从数学性能提升到临床效果，需开展循序渐进的实施和评估，解决其相互作用的复杂问题。

AI 系统的早期临床评价介于临床前开发（计算机模拟）、离线验证和临床试验之间，类似于外科领域 IDEAL 框架的 stage 1（设想）、2a（开发）或 2b（探索）[5,6]，或药物临床试验的 1 期与 2 期临床试验（图 1），关注 AI 系统的实际临床性能、安全性和人因学评价（human factors）。但目前少有 AI 系统临床研究涉及人因学评价，且可用性评估方法不一致。另外，也缺少 AI 系统运行环境、用户特征、选择过程及算法识别等方面的报告。

为提高 AI 系统临床研究的透明性和规范性，国际上已相继制订和发布相关报告规范，如：用于报告诊断或预后预测模型开发、验证和更新的 TRIPOD-AI[7]；用于报告诊断准确性研究的 STARD-AI[8]；分别用于报告评价随机对照试验及计划书的 CONSORT-AI[9] 和 SPIRIT-AI[10]（图 1），但 AI 系统早期临床评价阶段的报告规范仍存在空白。为改善实践中此类研究报告不充分的问题，2022 年 5 月，Vasey 等[11] 在 *BMJ* 发表了人工智能驱动下决策辅助系统早期临床评价的报告指南（reporting guideline for the developmental and exploratory clinical investigations of decision support systems driven by artificial intelligence，DECIDE-AI）。本文就指南的制订过程和主要内容进行解读，以期促进报告指南的正确理解和使用。

2　DECIDE-AI 的制订过程

DECIDE-AI 的制订参考 EQUATOR 协作网指南制订的基本流程进行，有专门的指导小组监督指南的制订过程[12]，见图 2：① 基于专家意见制订了初步的候选清单，该清单重点参考了基于人工智能诊断决策辅助系统的相关文献、指导小组成员推荐的文献及监管机构文件；② 通过不同的渠道招募专家，包括：指导小组推荐的专家、检获文献的作者、任何主动联系指导小组的专家及德尔菲专家推荐的专家（滚雪球），最终招募了行政人员/医院管理人员、医疗专业人员、临床医生、工程师/计算机科学家、人因设计专家、流行病学家、伦理学家、期刊编辑、患者代表等来自 18 个国家的 20 类利益相关群体参与；③ 开展两轮改良德尔菲专家咨询形成初步条目，138 名专家同意参加首轮德尔菲调查，其中 123 名（89%）完成了调查问卷，162 名专家受邀参加第二轮德尔菲调查，其中 138 人完成了问卷调查（85%）；④ 召开三轮专家线上共识会对初始条目进行增减、修改或补充，为确保关键利益相关群体的平衡及地域多样性，共识小组的 16 名专家参与讨论，最终确定 27 个条目；⑤ 将指南及解释性文件发给独立于共识小组之外的 16 名专家，确定最终报告清单中的条目及文字表述。

3　DECIDE-AI 主要内容

DECIDE-AI 报告指南由 17 条 AI 相关特异性报告条目（1 ~ 17）和 10 条一般性报告条目（I ~ X）组成，包括标题和摘要、引言、方法、结果、讨论和声明六部分（表 1），本文重点介绍与 AI 相关的特异性报告条目，条目中涉及的具体术语见表 2。

3.1　题目和摘要

题目中需明确说明研究是人工智能辅助决策系统的早期临床评价，帮助读者快速、准确的识别和检索研究。题目需体现：① 在辅助决策系统中使用机器学习/人工智能；② 辅助决策系统需解决

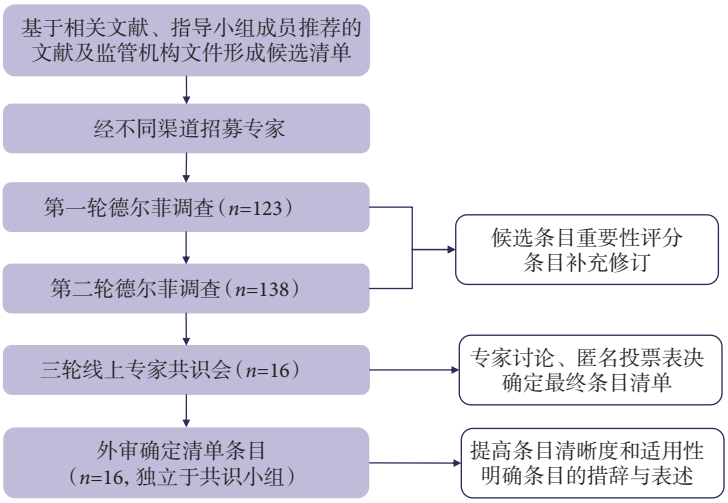

图 2　DECIDE-AI 制订过程流程图

的临床问题；③ 研究阶段。例如：基于人工智能的自主诊断系统在初级诊疗中检测糖尿病视网膜病变的关键性试验[19]。摘要部分建议使用结构式摘要，说明 AI 系统的预期用途、算法类型、研究环境、参与者、结局指标、安全性和人因学评价、主要结果和结论。

3.2　引言

引言部分要求说明 AI 系统的预期用途（或预期目的）及相关使用规范。这些信息与预期用途有关，不应与研究期间 AI 系统的实际使用情况相混淆，清晰描述预期用途有助于读者评估在相关场景中使用的 AI 系统是否代表预期用途。同时有助于监管机构参考临床研究中声称的预期用途决定新器械分类和审批。如果临床研究的预期用途与临床前开发阶段的预期用途不同，需明确说明。

预期用途部分要描述目标疾病/健康状况（如：败血症）和拟解决的问题（如：在液体和血管加压药剂量之间找到最佳平衡），明确定义当前针对此疾病/健康状况的标准实践方案及目标患者人群（条目 2a）。提供有关 AI 系统预期实施的信息，包括可能影响用户与 AI 系统交互的相关特征（如：用户在医疗保健系统中的角色和职责、专业、培训水平、对数字技术的熟悉程度等）、拟在临床路径的整合方式（使用环境、系统访问的难易程度、辅助决策的类型与时机等）及应用系统的潜在临床效果，对旨在改善患者医疗服务的 AI 系统，作者应说明针对哪些患者结局（如：30 天再入院率或死亡率）（条目 2b）。

3.3　方法

方法学部分重点报告以下六方面：

3.3.1　*参与者*　描述患者与数据层面的纳入排除标准（条目 3a）。患者层面的标准包括有关招募策略（在社区中主动、被动、开放招募）、抽样方法（连续、随机等）和知情同意（知情豁免）等信息。数据层面的标准包括采集时间、采集方法、数据质量、数据完整性和数据格式。需注意的是符合患者层面纳入标准的参与者可因数据质量低或数据不完整而被排除。样本量计算方面，AI 系统临床评价早期不要求正式统计样本量计算，但需说明事前确定的样本量。

与条目 3a 类似，作者还应详细描述用户的纳入排除标准及招募数量（3b），由于患者和用户都被视为参与者（见词汇表），还应报告用户获得知情同意的信息。但考虑到无论何种质量的用户数据均可提供人工智能系统可用性的信息（如：使用困难、缺乏兴趣等），通常不建议设置数据层面排除标准。

学习曲线评估是评价创新 AI 系统的重要内容之一，作者需详细说明为使用户熟悉 AI 系统而采取的措施，如培训课程类型、培训次数和时间等（条目 3c）。

3.3.2　*人工智能系统*　作者清晰描述 AI 系统的算法类型（即数学模型）、支持软硬件及其版本号，说明算法训练集中患者特征及系统在临床前开发或离线验证中的性能。如果上述信息无法从公开发表的文章中直接引用，建议以附件形式补充完整（条目 4a）。

详细描述输入数据特征，包括输入数据项清单、数据采集的时间范围、输入数据来源（如：常规收集数据、主动收集数据）、数据采集方式（如：计算机断层扫描仪成像、切片计数）、数据输入方式（如：从 EHR 中自动提取、手动输入等）、数据

表 1 DECIDE-AI 报告清单

条目	主题	具体内容
题目和摘要		
1	**题目**	**题目能识别该研究为基于人工智能或机器学习的辅助决策系统的早期临床评价，并阐明所解决的问题**
I	摘要	结构式摘要。包括：AI系统的预期用途、算法类型、研究场景、纳入患者和用户的数量、主要和次要结局、关键安全指标、人因学评价、主要结果及结论
引言		
2	**预期用途**	**a) 描述目标疾病/健康状况和拟解决的问题，包括目前标准实践方案和预期患者人群**
		b) 描述AI系统的预期用户、拟在临床路径中的整合方式及预期影响（包括患者结局）
II	目的	说明研究目的
方法		
III	研究相关信息	提供研究方案、研究注册号和伦理信息
3	**参与者**	**a) 描述患者招募情况，说明患者和数据层面的纳排标准，及如何决定患者招募数量**
		b) 描述用户招募情况，说明纳入和排除标准，及如何决定用户招募数量
		c) 说明为使用户熟悉AI系统而采取的措施，包括研究前接受的培训
4	**人工智能系统**	**a) 简要描述AI系统，包括算法类型及版本。描述或以参考文献的方式提供算法训练集中患者群体特征及系统在临床前开发（计算机模拟）/离线验证中的性能**
		b) 确定用作输入的数据，说明如何获取数据、数据输入过程、数据预处理及缺失/低质量数据的处理
		c) 描述AI系统的输出结果，及如何呈现给用户（可用图形展示）
5	**实施**	**a) 描述评价AI系统的环境**
		b) 描述评价AI系统的临床路径/临床工作流程、使用时机，及由谁和如何在AI支持下做出最终决策
IV	结局	明确测量的主要结局和次要结局
6	**安全性及错误**	**a) 说明如何定义和识别重大错误/故障**
		b) 描述如何识别、分析和最小化患者安全风险或伤害事件
7	**人因学**	**描述人因学使用的工具、方法或框架，设备的典型使用示例及如何选择参与人因评估的用户**
V	分析	描述主要结局和次要结局的统计分析方法和其他事前设定的分析，包括亚组分析，并说明理由
8	**伦理**	**描述是否使用特定方法来达到与伦理相关的目标（如算法公平性），并解释使用这些方法的理由**
VI	患者参与	说明患者如何参与以下工作，如：研究问题的提出、研究设计和研究实施
结果		
9	**参与者**	**a) 报告纳入患者的基线特征，并报告输入数据的缺失情况**
		b) 报告纳入用户的基线特征
10	**实施**	**a) 报告用户使用AI系统的情况，包括使用次数及是否按预期使用**
		b) 报告因使用AI系统引起的临床工作流程或临床路径的重大变化
VII	主要结果	报告预先设定的结局，包括对照组的结局（如果适用）
VIII	亚组分析	根据事前设定的亚组分析计划报告主要结果的亚组差异
11	**调整**	**报告研究期间对AI系统进行的任何调整，包括调整的时间、理由及调整后观察到的结果是否发生变化**
12	**人机协议**	**报告用户对AI系统的反馈，描述用户决策与AI系统建议不一致的情况及原因，如适用，还应当描述用户是否根据AI系统建议转变决策**
13	**安全性及错误**	**a) 列出与AI系统建议、支持软件/硬件及用户相关重大错误/故障，包括：① 发生率；② 原因；③ 是否可以纠正；④ 对患者服务的影响**
		b) 报告研究期间发现的患者安全风险或观察到的伤害事件（包括间接伤害）
14	**人因设计**	**a) 根据公认的标准或框架报告可用性评估结果**
		b) 报告用户学习曲线评估结果
讨论		
15	**支持预期用途**	**讨论研究结果是否支持预期临床场景下AI系统的预期用途**
16	**安全性及错误**	**讨论研究结果对AI系统安全的影响。包括任何观察到的错误/故障、伤害事件及其对患者医疗服务的影响，并讨论是否及如何可以减轻这些影响**
IX	优势和局限性	讨论研究的优势和局限性
声明		
17	**数据可获得性**	**说明能否及如何获取数据和相关代码**
X	利益冲突	说明任何相关的利益冲突，包括研究资金来源、资助者角色、商业公司参与情况，以及每位作者的个人利益冲突

1～17：与AI相关的特异性报告条目；I～X：一般性报告条目。

表 2　报告规范中相关术语定义

术语	定义
人工智能系统	结合人工智能的决策辅助系统，包括：① 人工智能或机器学习算法；② 配套软件平台；③ 配套硬件平台
人工智能系统版本	人工智能系统的形式及配套在某个时间点的唯一参考，可用于追踪人工智能系统随时间的变化，并对不同版本进行比较
算法	负责从数据中学习并生成输出的数学模型
人工智能	开发计算机系统，使其能执行通常需要人类智能完成任务的科学[13]
偏倚	与其他相比，在对待某些对象、人或群体方面存在的系统差异[14]
临床路径	患者在与医疗系统接触过程中经历的系列医疗程序和活动
临床相关的	与患者实际的观察和治疗相关，而不是计算机模拟或基于情景的模拟
临床评价	当人工智能系统按预期使用时，分析临床数据并使用科学方法来评估其在临床实践中的性能、有效性和/或安全性的系列活动[15]
临床研究	对一个或多个人类受试者进行的研究，以评估人工智能系统的临床表现、有效性和/或安全性[16]，可以在任何医疗场景下（例如，社区、初级保健、医院）进行
临床工作流程	医护人员为患者提供医疗服务时遵循的步骤和程序
辅助决策系统	通过提供针对个人和情境的具体信息或建议来支持人类做出决定，以改善服务或增进健康的系统
暴露	持续使用或曾经使用过人工智能系统或类似的数字技术
人机交互	通过物理接口和概念接口在人类用户和数字系统间产生的双向影响
人因学	也叫人体工程学，是研究人类与系统中其他要素间交互作用的科学学科，也是将理论、原则、数据和方法应用于设计，优化人类福祉和系统整体性能的专业
使用指征	使用人工智能系统的情况和原因（疾病/健康状况、相关临床问题和患者群体）
计算机生物模拟评价	通过计算机模拟技术在临床场景外开展的评价
预期用途	开发者声明的人工智能系统的预期用途，是人工智能系统监管分类的依据。包括：目标医疗环境、患者群体、用户群体、使用环境、作用方式
学习曲线	用户表现与经验的图形呈现[17]，用于分析用户任务性能随任务暴露增加而演变的过程
实际评价	真实临床条件下开展的评价，此环节产生的决定将直接影响患者服务；与“潜在”或“隐形”模式不同，后者的评价结果不会对患者产生直接影响
机器学习	涉及模型/算法开发的计算机科学领域，可通过对数据模式进行学习而非完全遵循设置好的规则来解决特定任务，是人工智能领域的一种方法[9]
参与者	参与研究的主体，即数据收集及知情同意（或知情同意豁免）的对象。DECIDE-AI 指南认为，患者和用户都可以是参与者
患者	接受医疗服务或使用医疗服务的个人，以及在人工智能系统支持下做出决定的主体
患者参与研究	“与”或“由”患者或公众一起进行的研究，而非“对”“关于”或“为”他们进行的研究
标准实践方案	特定医疗环境和问题下目标患者群体所接受的常规医疗服务，标准实践方案未必是最佳实践方案
可用性	在指定的环境中，特定用户可以使用产品有效、高效和满意的实现特定目标的程度[18]
用户	使用人工智能系统开展辅助决策的人，可以是医疗保健专业人员，也可以是患者

所给出的定义与DECIDE-AI的特定背景和指南中术语的使用有关。它们不一定是普遍认可的定义，也不一定完全适用于其他研究领域。

预处理及如何定义和处理缺失值（条目 4b）。

描述如何向用户呈现 AI 系统的输出结果，包括 AI 系统的输出类型和数量（如：AI 系统对每个检测到的结节进行分割并给出恶性肿瘤的概率），显示界面的设计（如：图像、屏幕截图、插图）及其他信息（如：关注机制的可视化，显示对 AI 系统推荐影响最大的变量等）。作者还应说明用户可对界面进行多大程度的定制，是否有机会让用户向 AI 系统提供交互反馈（条目 4c）。

3.3.3　*实施*　描述评价 AI 系统的环境，包括医疗中心的类型和规模（如：重大创伤中心），场所（如：急诊科）、相关人员和技术支持（如：多学科创伤团队、床旁射线照相术），或 AI 系统支持硬件（计算机）（条目 5a）。

描述研究期间如何使用人工智能系统的信息，包括与临床工作流程/临床路径集成相关的信息（如：患者的初始情况及其接受治疗的原因、使用 AI 系统做出的临床决策）及决策过程，包括涉及哪方面的人员、处于哪个阶段以及谁负责最终的临床决策（条目 5b）。

3.3.4　*安全性与错误*　说明如何明确定义和识别重大错误或故障，包括算法错误（如：错误的将结节描述为恶性）、支持软硬件故障（如：因数据提取或电池电量耗尽无法生成推荐意见）及涉及用户的错误（如：用户输入错误）（条目 6a）。

说明如何识别、分析和最小化患者安全风险或伤害事件，包括：伤害事件发生的可能性、对参与者的潜在影响、风险检测的难易程度及目标患者群

体的疾病严重程度（条目 6b）。

3.3.5 *人因学* 描述人因学使用的工具、方法或框架，设备的典型使用示例及如何选择参与人因评估的用户（条目 7）。与安全性一样，人因设计评估应在临床前阶段就已开展，这里主要指在临床实时环境下的持续评估。最合适的人因设计评估取决于环境和设备，主要评价其可用性。可用性评估需使用经过验证的工具、方法或框架，如：ISO 标准（ISO 16 982: 2002[20]；ISO 9 241-11: 2018[21]）、IEC 标准（IEC 62 366-1: 2015[22]；IEC/TR 62 366-2: 2016[23]；IEC 62 366-1:2015[24]），评估内容包括：任务完成时间、工作量、显示界面或用户满意度问卷。

3.3.6 *伦理* 描述是否使用特定方法来达到与伦理相关的目标（如算法公平性），并解释使用这些方法的理由。相关方法包括用于检测、量化和减轻算法输出中偏见的措施，包括但不限于算法公平性。例如，由于参考标准增加了黑人患者的估计风险，需重新调整心脏手术风险评估的算法（条目 8）。

3.4 结果

3.4.1 *参与者* 根据 AI 系统的预期用途、已知对结果有影响的因素选择要报告的基线特征（条目 9a）。例如：年龄、性别、种族、社会经济地位、地理位置、目标疾病的患病率、目标疾病的分类/严重程度、算法中包含的关键预测因子等。作者还应同时定量报告研究期间 AI 系统输入数据（条目 4b）的缺失情况，最好按数据项细分。

考虑报告用户的医学专业、培训水平、临床角色/资历、对决策的熟悉程度及他们之前是否接触过决策辅助工具等（条目 9b）。在用户数量较少的研究中，作者还需仔细考虑如何在报告用户基线特征时保持用户的匿名性。

3.4.2 *实施* 报告实际接触过辅助决策工具的潜在用户比例、有权访问该工具的用户使用该工具的频率、未能遵守 AI 系统指示使用的情况（如：适应症、使用时间等）（条目 10a）。如适用，还应简要描述本应使用 AI 系统但没有使用的情况。

报告 AI 系统对临床工作流程或临床路径造成的任何重大变化（条目 10b），注意区分临床工作流程（即医护人员为患者提供医疗服务时遵循的步骤和程序）和临床路径（即患者在与医疗系统接触过程中经历的系列医疗程序和活动），选择报告那些重大变化时应考虑：① 与条目 2b 描述的预期用途的区别（如：AI 系统原本旨在减少使用不适当的影像检查，但意外导致专科转诊数量的增加）；② 对患者安全的潜在风险；③ 对 AI 系统集成和接受程度的潜在影响。

3.4.3 *调整* 说明研究期间对 AI 系统做出的任何调整（条目 11），包括对算法的更改（如：重新校准）或对其支持硬件平台的更改（如：显示界面改进）等，详细记录更改后的版本号及这些更改对主要研究结果的影响。

3.4.4 *人机协议* 辅助决策系统旨在影响用户的决策，根据用户对人工智能系统建议的反应，可能会出现三大情况：① 决定/行动没有变化；② 决定/行动有所改善（凸显人工智能系统的潜在附加值）；③ 决定/行动恶化（使用人工智能系统会使患者面临额外的风险）。作者应详细报告用户对 AI 系统的反应，描述用户决策与 AI 系统建议不一致的情况及原因（条目 12）。如：初始用户决策、人工智能系统推荐、最终用户决策、临床情况、患者/病例特征、用户特征、改变的原因、改变的后果等。

3.4.5 *安全性及错误* 报告所有观察到的重大错误/故障（建议以表格形式列出），包括出现次数、原因、如何纠正相应错误/故障及对患者结局产生的影响（条目 13a）。作者应根据条目 6b 报告出现的患者安全风险或伤害事件（条目 13b）。

3.4.6 *人因学* 人因学评价结果的报告应以所选方法为指导（条目 7），如果与用户群体不同（或子集），则应指定人因设计评估参与者的特征（条目 14a）。统计描述有助于读者理解学习曲线的含义，图形的方式可为读者提供更精细的信息（条目 14b）。

3.5 讨论

3.5.1 *支持预期用途* 作者应根据结果描述对评估系统的实际预期，及这些结果如何支持系统的预期用途，并与当前标准实践方案和类似研究进行比较（条目 15）。结合人因设计评估结果讨论关键临床表现的结果，同时说明在采用 AI 系统开展下一阶段更大规模比较试验时可能存在的挑战。

3.5.2 *安全性及错误* 作者应结合错误/故障、已识别的风险、观察到的不良事件、临床路径的意外变化及与安全相关的人因技术评估结果总结该研究的主要安全发现，并提出可能的解决方案，如：算法再训练、产品再开发或修改后续试验设计等（条目 16）。

3.6 声明（数据可获得性）

说明能否公开获取算法和相关支持软件代码，如不能，需说明原因，若能，应说明获取途径（条目 17）。

4 小结

大数据时代下，AI 临床辅助决策系统在国内的应用愈发广泛，如新冠期间深圳大学医学部吴光耀教授团队开发的用于新冠肺炎患者入院时风险评估的辅助决策系统[25]；北京天坛医院李子孝教授团队开发的脑血管疾病 AI 临床辅助决策系统[26]；中国临床肿瘤学会开发的 CSSO 人工智能辅助决策系统等[27]。DECIDE-AI 指南经系统文献综述、专家咨询和国际专家共识会等步骤制订而成，综合各利益相关方建议，从 AI 辅助决策系统的预期用途、参与者、AI 算法、实施与应用、安全性与错误、人因学分析等方面为 AI 系统早期临床评价研究的报告提供强有力的指导，有助于提高 AI 系统早期临床评价研究报告的清晰度和透明度，并在研究设计、方案起草、研究注册中为研究者提供方法学支持，促进 AI 系统的临床应用。

需注意的是：① AI 系统的早期科学评估与监管在内容上存在一定程度的重合，但考虑到科学评估和监管评估的重点略有不同且国家间的监管策略存在差异[28]，该指南中未涉及监管内容，对监管的指导意义有限；② AI 系统输出结果的可解释性对提高用户和患者对 AI 系统的信任度及扩大实际应用范围至关重要[29]，但在共识过程中有专家认为 AI 系统的临床价值可能与可解释性无关，另外，由于目前尚缺少普遍接受的量化或评价可解释性的方法，共识小组最终决定在本指南中暂不列入与可解释性相关的条目；③ 随着用户积累 AI 系统的实际使用经验，其对 AI 系统推荐结果的信任度也会发生变化，了解信任度的变化趋势，有助于制订用户培训计划和确定比较试验中数据收集的最佳时间点，但与可解释性一样，因目前缺少达成共识的评价方法，本报告指南中暂未考虑信任度变化的问题。

DECIDE-AI 是不同专业背景、经验专家共识的结果，为人工智能辅助决策系统的早期临床评价提供最低报告标准，适用于该阶段所有研究设计类型和 AI 功能模式（检测、诊断、预后、治疗）的报告，但仍有局限性：① 虽然共识专家的选择以地域、专业多样化为原则，但仍存在地域（以欧洲为主）和利益相关群体（以临床医生和工程师为主）的不均衡，可能导致参与者选择偏倚；② 与其他 AI 系统报告规范类似，AI 系统早期临床评估的研究示例很少，可能影响从文献中提取初始条目的完整性，研究小组也是通过两轮德尔菲咨询尽量完善、补充条目；③ 目前该报告规范还处于实践初期，后期在真实世界更广泛群体中的应用将会促进条目的修改与完善。

综上，将人工智能引入医疗系统需得到完整、可靠及全面的证据支持，有助于确保人工智能系统的安全性和有效性及获得患者、从业者和购买者的信任。DECIDE-AI 旨在改善 AI 系统早期临床评价的报告，为后期更大规模的临床研究和广泛应用奠定基础。

参考文献

1 Liu X, Faes L, Kale AU, *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health, 2019, 1(6): e271-e297.

2 Vasey B, Ursprung S, Beddoe B, *et al*. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. JAMA Netw Open, 2021, 4(3): e211276.

3 Freeman K, Geppert J, Stinton C, *et al*. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ, 2021, 374: n1872.

4 Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med, 2018, 1: 40.

5 喻佳洁, 陕飞, McCulloch P, 等. 外科创新技术/器械临床研究方法学—IDEAL 框架与推荐系列文章之一: IDEAL 框架与推荐介绍. 中国胸心血管外科临床杂志, 2021, 28(2): 131-136.

6 喻佳洁, Hirst A, McCulloch P, 等. 外科创新技术/器械临床研究方法学—IDEAL 框架与推荐系列文章之二: IDEAL 报告规范解读. 中国胸心血管外科临床杂志, 2021, 28(3): 263-270.

7 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet, 2019, 393(10181): 1577-1579.

8 Sounderajah V, Ashrafian H, Aggarwal R, *et al*. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. Nat Med, 2020, 26(6): 807-808.

9 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health, 2020, 2(10): e537-e548.

10 Cruz RS, Liu X, Chan AW, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med, 2020, 26(9): 1351-1363.

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. BMJ, 2022, 377: e070904.

12 Moher D, Schulz KF, Simera I, *et al*. Guidance for developers of health research reporting guidelines. PLoS Med, 2010, 7(2): e1000217.

13 Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med, 2019, 11(1): 70.

14 International Organization for Standardization. Information technology - artificial intelligence (AI) - bias in AI systems and AI aided decision making (ISO/IEC TR 24027:2021). 2021.

15 US Food and Drug Administration (FDA). Clinical decision

support software: draft guidance for industry and Food and Drug Administration staff. 2019.
16 Vasey B, Nagendran M, Mcculloch P. DECIDE-AI 2022. 2022.
17 Hopper AN, Jamison MH, Lewis WG. Learning curves in surgical practice. Postgrad Med J, 2007, 83(986): 777-779.
18 Bilbro NA, Hirst A, Paez A, *et al*. The IDEAL reporting guidelines: a Delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. Ann Surg, 2021, 273(1): 82-85.
19 Abràmoff MD, Lavin PT, Birch M, *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med, 2018, 1: 39.
20 International Organization for Standardization. Ergonomics of human-system interaction-usability methods supporting human-centred design. ISO/TR 16982, 2002.
21 International Organization for Standardization. Ergonomics of human-system interaction-part 11: usability: definitions and concepts. ISO 9241-11, 2018.
22 International Electrotechnical Commission. Medical devices-part 1: application of usability engineering to medical devices. IEC 62366-1, 2015.
23 International Electrotechnical Commission. Medical devices-part 2: guidance on the application of usability engineering to medical devices. IEC TR 62366-2, 2016.
24 International Electrotechnical Commission. Medical devices-part 2: guidance on the application of usability engineering to medical devices-amendment 1. IEC 62366-1, 2015.
25 Wu G, Yang P, Xie Y, *et al*. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. Eur Respir J, 2020, 56(2): 2001104.
26 Li Z, Zhang X, Ding L, *et al*. Rationale and design of the GOLDEN BRIDGE II: a cluster-randomised multifaceted intervention trial of an artificial intelligence-based cerebrovascular disease clinical decision support system to improve stroke outcomes and care quality in China. Stroke Vasc Neurol, 2024.
27 李健斌, 江泽飞. 中国临床肿瘤学会人工智能决策系统的建立与应用. 中华医学杂志, 2020, 100(6): 411-415.
28 Boel A, Navarro-Compán V, Landewé R, *et al*. Two different invitation approaches for consecutive rounds of a Delphi survey led to comparable final outcome. J Clin Epidemiol, 2021, 129: 31-39.
29 Sujan M, Furniss D, Grundy K, *et al*. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inform, 2019, 26(1): e100081.

收稿日期：2024-01-30 修回日期：2024-04-07
本文编辑：蔡羽嘉