TC260-PG-2026NA

网络安全标准实践指南

——人工智能应用安全指引 卫生健康

（征求意见稿 v1.0-202606）

全国网络安全标准化技术委员会秘书处

2026 年 06 月

本文档可从以下网址获得：
www.tc260.org.cn/

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

前　言

《网络安全标准实践指南》（以下简称《实践指南》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织制定和发布的标准相关技术文件，旨在围绕网络安全法律法规政策、标准、网络安全热点和事件等主题，宣传网络安全相关标准及知识，提供标准化实践指引。

本文件起草单位：国家卫生健康委统计信息中心、中国医学科学院阜外医院、中国电子技术标准化研究院、北京市卫生健康大数据与政策研究中心、华中科技大学同济医学院附属同济医院。

本文件主要起草人：赵韡、李岳峰、韩作为、郑攀、任宇飞、张妍婷、张黎黎、张宇希、汪洋、庹兵兵、吕飞霄。

声　明

本《实践指南》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《实践指南》的任何部分。凡转载或引用本《实践指南》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。

摘　要

为全面提升各行业人工智能应用安全水平，以高水平安全保障人工智能高质量发展，依据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》，以及《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等政策法规，制定人工智能应用安全指引系列文件。

人工智能应用安全指引系列文件包含通用文件以及行业领域文件两类。通用文件提供通用性实践安全指导，适用于各行业领域开展人工智能应用活动。行业领域文件给出适用于特定行业领域的实践指引。

本《实践指南》属于行业领域文件，用于指导各级卫生健康行政部门、各级各类医疗卫生机构和相关事业单位等卫生健康机构及技术提供方开展人工智能应用活动。本文件与《网络安全标准实践指南——人工智能应用安全指引 总则》等通用文件配合使用。

目 录

1 范围

本文件给出了卫生健康领域应用人工智能的安全指引，包括总体业务安全、应用流程安全、应用场景安全等方面。

本标准适用于指导各级卫生健康行政部门、各级各类医疗卫生机构和相关事业单位等卫生健康机构及技术提供方在卫生健康领域开展人工智能应用活动。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本标准必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本标准；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本标准。

GB/T 35273 信息安全技术 个人信息安全规范

GB/T 41867 信息技术 人工智能 术语

GB/T 45438 网络安全技术 人工智能生成合成内容标识方法

GB/T 45654 网络安全技术 生成式人工智能服务安全基本要求

GB/T 45674 网络安全技术 生成式人工智能数据标注安全规范

YY/T 0664 医疗器械软件 软件生存周期过程

TC260-PG-2026XX 网络安全标准实践指南——人工智能应用安全指引 总则

TC260-005 人工智能应用伦理安全指引 1.0

3 术语和定义

TC260-PG-2026XX《网络安全标准实践指南——人工智能应用安全指引 总则》附录 B 中界定的术语和定义适用于本标准。

4 概述

人工智能技术在卫生健康领域的应用涉及应用场景、应用流程、风险等级等方面，是一个多维、动态、场景化的复杂系统，应依据其所属的应用场景特性、所处的生命周期阶段以及被评定的风险等级进行综合评估，实施差异化、精细化的安全管理策略。

卫生健康机构在开展人工智能应用活动时，应符合国家相关法律法规和政策文件的要求，满足行业发展需求，在遵循 TC260-PG-2026XX《网络安全标准实践指南——人工智能应用安全指引 总则》基础上，进一步按照本文件第 5 章至第 7 章提出的安全指引，防范人工智能应用安全风险，保障人工智能应用安全可控。

总体设计内容包括:

X-应用场景：如医疗辅助支持、患者服务与应用、公共卫生与应急、健康教育与促进、卫生管理与其他。

Y-应用流程：人工智能应用的安全贯穿其从设计到下线的完整生命周期，包含人工智能应用的前期规划、设计开发、验证确认、部署、

运行和监控、持续验证评估、退役下线等过程。

注：应用流程参考 TC260-PG-2026XX《网络安全标准实践指南——人工智能应用安全指引 总则》第 6 章。

Z-风险等级：包含特别重大风险、重大风险、较大风险、一般风险、低风险五类风险等级及对应的安全要求。

总体设计见图 1。

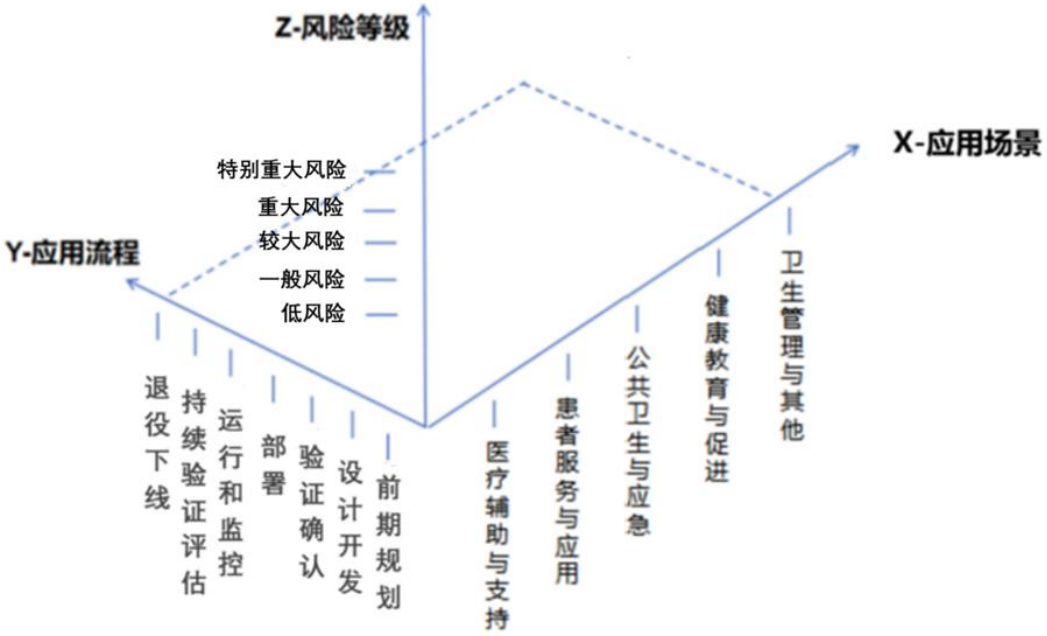

图1 总体架构图

5 总体原则

——保护人类的自主性。医疗人工智能系统的部署应始终维护人类在医疗决策中的核心地位，确保医护人员享有最终的临床控制权，并充分保障患者在知情同意基础上的自主选择权。

——确保透明性和可解释性。医疗人工智能应在开发与应用阶段公开其数据来源、算法逻辑及技术局限性，以可理解的方式向医患双

方解释其决策依据，从而建立合理的信任。

——确保包容性和公平性。医疗人工智能应积极消除算法和数据集中的潜在偏见，确保技术能够公平地惠及不同民族、性别、地域和经济背景的群体，避免加剧现有的健康不平等。

——风险分级分类。应根据人工智能应用在不同医疗场景和不同流程阶段中对患者生命安全的潜在影响程度，实施差异化的风险分级分类管理，在鼓励技术创新的同时坚守安全底线。

——培养责任感和实行问责制。应建立清晰的伦理与法律问责框架，明确开发者、医疗机构及监管者的权责边界，确保当人工智能系统发生错误或引发不良后果时，有明确的追责与救济机制。

——推广反应迅速且可持续的人工智能。宜具备敏捷适应临床需求和公共卫生环境动态变化的能力，同时其开发与运行应注重资源效益，尽量减少对环境的负面影响，以实现技术的长期可持续发展。

——增进人类福祉、安全和公共利益。人工智能的设计与应用应严格符合安全性、有效性和质量控制的要求，提高医疗服务水平、改善患者健康结果并促进全球公共卫生利益。

——强化数据安全与隐私保护。医疗人工智能涉及的患者个人健康信息、诊疗数据等，应严格遵循隐私保护相关法律法规。明确数据采集、存储、传输、使用、销毁的安全规范，采取加密、脱敏等技术措施，严防数据泄露、滥用、篡改，保障患者隐私与数据安全，是所有原则落地的基础前提。

注：参考 TC260-005《人工智能应用伦理安全指引 1.0》中第 5 章。

6 业务安全指引

6.1 医生主导与人工智能辅助

医疗行为的决策主体由医生担任，人工智能仅作为辅助工具，不得替代医生的专业判断及伦理责任。主要包括:

a) 临床决策权限：人工智能输出的诊断建议、治疗方案等结果需经医生审核确认，医生应结合患者病史、生活方式等个体情况综合判断，不得盲目依赖人工智能系统。

b) 人机协作机制：人工智能系统设计应具备医生交互功能，例如提供差异化建议的可视化对比、不确定性提示，告知医生认知人工智能的技术局限性。

c) 责任划分规则：明确医疗事故中医生与人工智能的责任边界。若人工智能提供错误建议且医生未识别，医生承担主要责任；若因人工智能系统设计缺陷导致事故，由卫生健康机构先行承担责任，机构有权依法向技术提供方追偿。

d) 医生培训要求：医生需接受人工智能工具使用培训，内容包括掌握适用范围和限制条件、识别潜在偏差（如特定人群的数据偏差）和执行紧急情况手动干预流程等。

6.2 患者知情与伦理合规

人工智能在卫生健康领域应用中，应保证患者知情同意及伦理合规要求。主要包括:

a) 知情同意机制：患者需知晓人工智能在诊断、治疗中的参与程度，并有权拒绝人工智能辅助。

b) 知情同意书内容：应包含人工智能技术用途、潜在风险、数据使用范围、患者权利等信息，确保语言通俗易懂，并明确告知人工智能生成内容。

c) 伦理审查要求：医疗人工智能应用需通过卫生健康机构伦理委员会审查，符合《涉及人的生物医学研究伦理审查办法》等规范，重点评估算法公平性及患者权益保护措施。

d) 伦理监督与反馈：建立患者投诉渠道，定期开展伦理评估，及时调整人工智能应用中可能违反伦理原则的环节。

7 应用流程安全指引

卫生健康机构在开展人工智能应用活动各流程阶段中安全指引包括：

a) 应参考 TC260-PG-2026XX《网络安全标准实践指南——人工智能应用安全指引 总则》，提升人工智能应用安全水平，防范人工智能应用安全风险。

b) 模型输出的结论应具备一定的可解释性，能够向医生展示其判断的关键特征和逻辑依据。对于高风险的决策，建立多重决策机制，最终诊断和治疗方案的决定权应掌握在合格的医生手中。

c) 应评估模型在不同人群亚组中的性能表现，确保其决策不会对特定群体产生系统性的不利影响。对于在特定人群（如儿童、老年人）中性能显著下降的模型，应限制其在该特定人群中的使用权限或明确提示置信度降低风险。

d) 应完整记录所有用户的操作日志、数据访问与输入输出日志、模型调用日志和系统异常等日志，满足医疗纠纷溯源与审计要求，不定期开展模型安全性测评分析。应对不良事件进行分析总结，对于高风险安全事件应立即进行整改。

e) 应通过伦理部门的审查，完成卫生健康行政管理部门备案，确保其符合医学伦理，保障患者权益。

8 应用场景安全指引

8.1 医疗辅助与支持

应用人工智能的医疗辅助支持，包括应急救护与远程医疗、生成式电子病历、影像辅助判读、临床决策支持、中医辅助决策、智能手术操作规划、院内智能用药管理、院内智能患者管理与功能评价、智能质控管理等。安全指引包括:

a) 技术提供方应确保网络与数据链路稳定并配置高容错预案与人工备份方案，医生应对关键生命体征与影像提示实施实时校验与人工复核，卫生健康机构应统一数据采集标准与阈值配置以控制误报与漏报并保持跨设备一致性。

b) 卫生健康机构宜加强数据隐私安全保护，以居民电子健康档案为核心建立患者数据调阅审批与留痕审计机制，落实分级授权、最小必要访问、异常访问告警与定期复核。

c) 技术提供方应明确算法适用范围与责任边界，提供循证依据与可解释信息。系统应由具有相应资质与经验的医生最终审核签署并通过多学科讨论优化方案，涉及诊疗、干预建议的采纳须由医生与患者共同确认，并以书面或电子知情同意形式留存，提供撤回同意与再沟通流程。

d) 卫生健康机构宜强化人员训练以提升在高压情境下的独立判断能力并建立预警闭环与抽查复核机制，治疗诊断中有涉及相关器械应提供相关备案并在使用前完成审查确认。

e) 技术提供方应对低置信度情形启用强制人工复核，宜定期收集医生反馈对模型输出、规则、阈值等进行校正。

8.2 患者服务与应用

应用人工智能的患者服务，包括智能问答、患者随访、慢病管理、智能客服、用药指导、适老化服务等。安全指引包括:

a) 对于智能预问诊与分诊导诊场景，技术提供方应确保医学术语映射的准确性，防止因理解歧义导致推荐错误的科室或遗漏危急重症症状。系统必须设置危急值识别与阻断机制，当识别到患者描述涉及急救指征时，应立即停止 AI 问答，弹出醒目警

示并引导至急诊或人工服务。必须明确导诊建议仅供参考，不作为最终确诊依据。

b) 对于患者随访与慢病管理场景，应由卫生健康机构发起，医生依据临床诊断和医嘱设定个体化干预方案，完成知情同意与状态指示配置，并与紧急联系人或专业服务中心建立联动形成闭环处置；卫生健康机构应在保障数据安全与隐私保护的前提下提供可穿戴装置的正确佩戴与使用指引，明确异常触发条件与提示方式。技术提供方宜采用人性化交互，提升依从性并建立用户反馈与责任追溯机制。

c) 对于智能客服与用药咨询场景，技术提供方应构建基于权威医学指南与药品说明书的专用知识库，宜采用相应技术确保回复内容的可溯源性，严禁系统生成无医学依据的建议。卫生健康机构应建立危急重症诱导词库与阻断机制，当识别到胸痛、呼吸困难等急救指征时，应强制中断 AI 对话并立即引导至人工服务或急诊通道；涉及处方药咨询与用药调整建议时，系统应显著提示“仅供参考，不作为最终处方依据”，相关建议须经具有执业资格的药师复核确认后方可推送。

d) 对于重点人群（一老一小）服务场景，卫生健康机构应建立监护人授权与账户关联机制，提供适老化与适幼化交互界面，宜支持语音、手势等多模态输入以降低使用门槛。技术提供方应针对老年人方言、口音及儿童语言特征进行模型专项训练与优

化，确保意图识别准确性；对于涉及跌倒监测、生命体征异常或生长发育评估等高风险提示，应设置双重验证逻辑，并建立向监护人及医疗团队同步预警的闭环处置流程。

8.3 公共卫生与应急

应用人工智能的公共卫生与应急，包括传染病智能监测、流行病学调查分析、突发公共卫生事件预警、卫生应急管理等。安全指引包括:

a) 卫生健康机构和技术提供方应落实数据安全与隐私保护、访问控制与留痕审计，应建立数据分级与访问审批机制，应对敏感数据实施脱敏或匿名化处理。

b) 技术提供方应建立反馈与纠错闭环，开展模型校准、效果评估与版本管理。宜整合疾控、急救、血液、监督等卫生健康信息数据资源。

c) 卫生健康机构宜建立多部门协同决策记录机制，明确职责分工与责任追溯流程，提升预警与处置的一致性。

d) 卫生健康机构宜按地区与人群分层输出告警与处置建议，设置阈值动态调整机制；针对群体性不明原因疾病等异常公卫事件，应强制人工介入，对人工智能结果进行交叉验证。

8.4 健康教育与科研

应用人工智能的健康教育与促进，包括智能药械研发与科教管理、智能健康风险评估与筛查、医学知识科普、多中心临床研究、药物研

发数据分析等。科研应落实研究中所涉及的个人信息安全、多中心数据安全和数据传输交换等安全保障措施。安全指引包括:

a) 卫生健康机构对提供的知识应准确权威来源，清晰表述规范并配套适用范围与模型局限说明，宜对健康教育类信息实施分级管理，区分风险相关与一般性内容，在涉及高风险情形时设置显著提示与专业转介路径。

b) 卫生健康机构应建立来源可靠且可追溯的知识库与数据资源校核与更新机制，并对涉及特殊人群的内容开展专家审核，应建立争议知识准入与拦截机制。技术提供方宜在小规模验证后组织推广并建立用户反馈与纠错通道，宜制定知识库与算法的定期评估与版本发布计划做好知识的更新。

c) 人工智能应用于药物临床试验设计、受试者招募、数据统计分析、安全性监测等环节时，卫生健康机构及技术提供方应明确人工智能与研究人员的权责边界，确保研究人员对临床试验的核心决策拥有最终控制权。研究项目应经伦理委员会审核、监管部门备案，明确方案的科学性、合规性与安全性。临床试验数据的人工智能分析结果，应经专业研究人员复核确认。

8.5 卫生管理与其他

应用人工智能的卫生管理与其他，包括医院管理、区域管理、医疗资源调度、医保基金监管、供应链管理、其他。安全指引包括:

a) 应用人工智能的医院管理，包括医院运营管理、后勤管理、设备管理等。技术提供方应重点保障业务连续性，系统应设计为无单点故障的冗余架构。应定义严格的恢复时间目标和恢复点目标，确保在主系统故障时能秒级或分钟级切换到备用系统。应备有详细的手动操作预案。当人工智能系统不可用时，应能迅速切换回传统的人工调度模式，并进行定期演练。应注意数据脱敏与数据安全，所有用于决策的关键数据流，应在传输过程中使用加密来保证机密性和完整性。

b) 应用人工智能的区域管理，包括区域协同管理、卫生资源规划与监管、卫生政务管理等。技术提供方应避免数据偏见导致宏观决策失当、数据被污染导致错误的资源配置和大规模隐私泄露等问题。关注数据隐私保护，系统应采用隐私增强技术，确保在不暴露个体信息的前提下，获得有价值的统计结果。在部署任何用于资源分配或监管的人工智能模型前，技术提供方应进行算法影响评估，向公众或受影响方解释模型的工作原理、数据来源和潜在的社会伦理影响，并接受质询。

附 录 A

（资料性）

卫生健康领域人工智能应用风险分级

A.1 风险识别

风险识别应结合系统功能、应用场景、数据性质、决策影响等因素进行综合分析，包含以下四个方面：

系统功能与决策介入程度：依据人工智能临床应用场景对辅助诊断、治疗规划、预后预测等不同临床决策的介入程度进行分类，判断系统是否属于“具有较高风险”的诊疗支持类技术。

数据敏感性与隐私影响：依据健康医疗数据的分级分类保护要求，识别系统所涉及的个人信息、病历数据、群体健康信息等是否属于核心或重要数据，并评估其泄露、篡改、滥用的可能性与影响。

设施潜在危害类型与严重性评估：结合技术风险预警与控制的要求，评估其错误输出可能导致的人身伤害、诊疗延误等后果。

伦理与社会影响：识别系统在公平性、可解释性、知情同意和算法歧视等方面可能引发的伦理与社会风险。

A.2 风险评估

卫生健康机构和技术提供方应在风险识别基础上，组织临床、技术、伦理等多学科团队从人工介入程度、结果严重程度、经济与社会影响等级、伦理与合规风险等级四个维度进行量化分析。评估过程应

形成书面记录，并纳入系统风险管理文档。

注：参考《人工智能安全治理框架》2.0 进行风险分级。

A.3 风险定级

a) 特别重大风险

特别重大风险人工智能系统具有灾难性和系统性威胁特征，其输出结果直接驱动、执行高风险诊疗行为，或在无人干预的情况下对群体患者、公共卫生安全产生重大生理影响的系统。具体包括：

人工介入程度：高度自主运行，人工仅在系统发生重大失效、引发严重风险时紧急介入终止。

结果严重程度：错误或失效可能导致患者死亡、不可逆的身体伤害或公共卫生危机，影响范围极广。

经济与社会影响等级：医疗辅助支持和决策等引发社会舆论、大规模数据泄露事件。

伦理与合规风险等级：涉及极高敏感人群或核心敏感场景，存在严重偏倚、重大隐私侵害风险，属于国家重点监管目录核心范围。

b) 重大风险

重大风险人工智能系统具有重大威胁性和区域性影响特征，其输出结果直接驱动或执行关键诊疗行为，或在无人干预的情况下持续对个别患者产生重大生理影响的系统。具体包括：

人工介入程度：人工介入能力受限，无法实时干预系统核心运行流程。

结果严重程度：错误或失效可能导致患者严重健康损害。

经济与社会影响等级：诱发较大规模数据安全事件，引发广泛社会关注。

伦理与合规风险等级：涉及高度敏感人群或场景，存在显著偏倚或隐私侵害风险，处于严格监管目录范围，需通过专项伦理审查。

c) 较大风险

较大风险人工智能系统具有明显威胁性和局部性影响特征，其输出结果对临床诊疗或个人健康决策有辅助作用，但最终决策仍需人类综合判断完成的系统。具体包括：

人工介入程度：人工审核介入，但可能面临警报疲劳、自动化偏见或认知负荷等挑战。

结果严重程度：错误可能导致患者可逆的身体伤害、病情延误、半敏感个人信息泄露等中度健康损害。

经济与社会影响等级：可能造成机构较大经济损失，引发局部关注。

伦理与合规风险等级：存在公平性、偏倚或隐私暴露等中等风险，需满足算法可解释性、公平性要求。

d) 一般风险

一般风险人工智能系统具有一定威胁性但影响范围有限，主要用于提升临床辅助信息处理效率或辅助非核心诊疗管理流程，其输出结果可间接影响基础健康决策。具体包括：

人工介入程度：人工主导系统运行，系统输出仅作为基础参考，不影响人工最终判断。

结果严重程度：错误可能间接造成轻微病情延误、基础个人信息泄露等影响，但可快速纠正。

经济与社会影响等级：仅造成轻微经济成本增加或工作效率波动，引发小范围关注。

伦理与合规风险等级：伦理风险低，需符合一般性医疗数据合规、隐私保护要求。

e) 低风险

低风险人工智能系统具有轻微威胁性且影响范围很小，主要用于提升非诊疗类信息处理效率、辅助行政办公或提供一般性健康科普知识，其输出结果与诊疗决策、个人健康无直接关联。具体包括：

人工介入程度：人工完全主导系统运行，拥有绝对控制与审核权限。

结果严重程度：错误不会对患者身体、健康产生任何影响，仅可能影响行政办公效率。

经济与社会影响等级：无明显经济损失，不影响关键业务连续性。

伦理与合规风险等级：无明显伦理风险，不涉及患者隐私、公平性等敏感问题。

参考文献

[1]中华人民共和国网络安全法

[1.1] GB/T 45438 网络安全技术 人工智能生成合成内容标识方法

[1.2] GB/T 45654 网络安全技术 生成式人工智能服务安全基本要求

[1.3] GB/T 45674 网络安全技术 生成式人工智能数据标注安全规范

[2] 中华人民共和国数据安全法

[3] 中华人民共和国个人信息保护法

[3.1] GB/T 35273 信息安全技术 个人信息安全规范

[4] GB/T 41867 信息技术 人工智能 术语

[5] 网络数据安全管理条例 国务院令 2024 年 9 月

[6] YY/T 0664 医疗器械软件 软件生存周期过程

[7] 人工智能安全治理框架 2.0 2025 年 9 月

[8] 人工智能应用伦理安全指引 1.0 2026 年 5 月