附件

# 人工智能医学影像研究伦理指引

## 1. 目的及范围

人工智能医学影像作为人工智能赋能医学研究和临床实践的重要组成部分，正在逐渐走向应用。为指导人工智能医学影像研究的合规开展，规范人工智能医学影像相关的基础算法研究、技术研发、临床验证和转化，防范科技伦理风险，推动人工智能医学影像研究的负责任创新与健康有序发展，特提出人工智能医学影像研究伦理指引。

## 2. 术语

2.1 人工智能医学影像（Artificial Intelligence in Medical Imaging）

人工智能医学影像是指利用人工智能技术对医学影像数据进行分析和处理的一种技术系统与方法，主要用于辅助医生开展疾病诊断、治疗决策、健康管理以及为医患共同决策提供信息支持。这些技术能够自动识别和解析医学影像，如数字放射摄影（DR）、计算机断层成像（CT）、磁共振成像（MRI）、单光子发射计算机断层成像（SPECT）、正电子发射计算机断层成像（PET）及超声（US）等的特征，提高诊断的准确性和效率。

2.2 人工智能医学影像数据（Artificial Intelligence Medical Imaging Related Data）

人工智能医学影像数据是指与人工智能医学影像研究相关的各类数据，按数据来源及用途分为:

（1）原始影像数据：探测器接收信号经放大和模数转换后形成的基础影像数据。

（2）训练数据：用于训练机器学习模型的输入数据样本子集。

（3）预训练数据：用于对模型进行初步训练，以学习通用特征表示的数据。这些数据包括：①领域内数据：来自医学影像领域但任务或标注不同的数据(如使用自然图像预训练的模型在医学图像上进行微调)；②领域外数据：来自非医学影像领域的大规模数据集(如使用自然图像数据集进行预训练)。使用预训练数据旨在提升模型在目标任务上的性能、收敛速度或小数据场景下的泛化能力。

（4）合成数据：通过数学模型、算法或生成式 AI 模拟生成的、非直接源于原始采集的影像数据，用于补充训练样本或数据增强。

（5）验证与测试数据：用于评估模型性能的数据集，分别承担参数优化（验证数据）与性能盲测（测试数据）的功能。

2.3 脱敏处理（Data Masking）

脱敏处理是指通过一系列数据处理方法对原始数据进行处

理，以屏蔽敏感数据的一种数据保护方式。

2.4 算法泛化能力（Algorithm Generalization Ability）

算法泛化能力是指机器学习模型在有限训练数据上学习后，能在来自同一或相近分布、但从未见过的新样本上仍保持较好预测或决策表现的能力。

2.5 算法黑箱（Algorithm Black Box）

算法黑箱是指内部结构不明，无法通过对其功能分析进行认识，也无法通过对其要素特性和要素之间的联系推断其特性的系统。关于系统行为的某些规律只能通过检查其对环境的输出以及这种输出对于输入的种种关系，即通过输入—输出关系来建立。

2.6 人工智能幻觉（AI Hallucination）

人工智能幻觉是指人工智能系统（含多模态医学影像模型、大语言模型）输出看似可信、形式完整，但与客观事实、原始数据源、输入条件不符或完全虚构内容的现象。

2.7 算法可解释性（Algorithm Explainability）

算法可解释性是指人工智能系统具备的一种能力或特征，它使特定人类受众能够理解系统行为的原因。

2.8 特异性（Specificity）

特异性是指被算法判为阴性的阴性样本占全体阴性样本的比例。

2.9 灵敏性（Sensitivity）

灵敏性是指被算法判为阳性的阳性样本占全体阳性样本的

比例。

2.10 精确率（Precision）

精确率是指所有预测为阳性的样本中，被算法正确判断为阳性的比例。

**3. 基本原则**

3.1 增进人类福祉

人工智能医学影像研究应以增进人类福祉为根本宗旨，尊重和保障生命权利，将患者健康利益置于优先位置。应严格恪守医学伦理底线，审慎研判技术研发与应用对人体健康、生命安全的深远影响，推动负责任的医学科技创新，使技术进步服务于人类健康，促进社会可持续发展。

3.2 促进公平公正

人工智能医学影像研究应坚持公平公正原则，维护技术成果的公共属性与全民可及性。关注和积极解决算法歧视、群体偏差问题，优化算法适配性与技术适配场景，提升技术包容性与普惠性。应重点关注儿童、老年人、罕见病患者等脆弱群体的差异化健康需求，促进医疗资源的公平分配，确保所有人平等享有健康权益。

3.3 尊重人的主体性

人工智能医学影像研究应尊重人的主体性，确立人工智能在决策中的辅助地位，确保人的决策权。最终的诊断和治疗方案必须由医生结合专业知识和患者的具体情况来制定，确保人类对医

疗行为拥有最终的控制权和责任。患者有权了解人工智能医学影像技术在其诊疗过程中的作用、潜在风险和局限性。研究应确保患者的知情同意权，使其能够自主决定是否接受人工智能辅助的诊断或治疗。

3.4 保护隐私和数据安全

人工智能医学影像研究应切实保护个人隐私、强化数据安全。对医学影像及健康数据的采集、传输、存储、分析、共享、销毁，应实施全流程规范管理，建立完善的数据授权与撤回机制。尤其是大规模数据存储应采取高等级的安全保障，防止隐私泄露、外部攻击和未经授权使用。

3.5 确保安全可控

人工智能医学影像研究应遵循医学科学规律，建立全生命周期风险治理机制，保障技术研发、实验与临床应用的安全、规范、有效、可控。健全风险事前预防、事中管控、事后处置的全流程闭环管理，守住医疗安全与科技伦理底线。

3.6 强化透明可信

人工智能医学影像研究应坚持透明可信原则，致力于提升系统的可解释性与可靠性，确保技术逻辑与决策机制可理解与验证。应建立涵盖研发、训练至应用部署的全生命周期可追溯管理机制。

**4. 一般要求**

4.1 合法合规

开展人工智能医学影像研究须符合我国相关法律法规，遵循

国际公认的伦理准则，以及科学共同体达成的专业共识和技术规范。从数据获取到算法研发、产品注册和临床部署，每一环节均应在法律法规、伦理规范的框架下进行。不得通过人工智能医学影像研究进行非法活动、侵害他人合法权益、破坏社会稳定。

4.2 知情同意

开展人工智能医学影像研究，应获得研究参与者的书面知情同意。知情同意书和知情同意过程应规范，并获得伦理审查委员会批准，必要时采用动态知情同意。应允许研究参与者撤销同意。对于数据提供者已经签署了知情同意书且已匿名化处理的回顾性医学影像数据，经伦理委员会审查确认符合《涉及人的生命科学和医学研究伦理审查办法》有关要求的，可免除知情同意。

4.3 责任机制

人工智能医学影像研究应建立以促进创新和保障安全为目标的协同责任机制。必须坚持人是技术设计、部署与应用的最终责任主体，明确科研人员、医疗机构、临床医生、科技企业和监管部门在全生命周期中各自的职责，确保人工智能医学影像研究结果可回溯、可审计、可问责。

4.4 利益冲突

人工智能医学影像研究过程中应披露所有可能影响研究客观性、公正性的利益关系。研究团队需建立利益冲突审查与管理机制。应对潜在利益冲突进行评估，并制定规避措施，以保障研究设计、数据采集、结果分析及成果转化的全过程不受利益干扰，

维护研究的科学性与伦理公正性。

**5. 特殊要求**

5.1 数据方面

在涉及人工智能医学影像的研究中，应重点关注数据全生命周期中的数据安全、数据质量和数据脱敏。

（1）数据安全。应建立覆盖人工智能医学影像研究全流程（含基础实验、临床试验、转化应用）的数据全生命周期安全管理体系，确保从采集、传输、存储、分析、共享到销毁各环节均符合国家法律法规，防止数据泄露、损毁、篡改或丢失。应制定科学、详尽且可执行的数据使用协议，涵盖数据用途、权限管理、访问控制、加密存储、备份恢复等要求，并建立监测与审计机制，确保数据使用过程可追溯、可验证。

（2）数据质量。应确保数据标注的专业性和准确性。对于标注者资质、图像征象、图像标注、图像分割以及图像量化方法，应建立并遵照统一标准。应使用来源清晰并能够被证实具有较好科学性的数据集，考虑所涉及的数据源为回顾性数据还是前瞻性数据，考虑数据来源分布、数据集的潜在偏倚、数据结构以及数据标注是否具备科学性。

（3）数据脱敏。因履行公共卫生统计等公共利益职责，或基于科学研究确需向技术研究团队或第三方共享相关数据时，对影像数据中涉及的医疗健康、生物特征等敏感个人信息，应采取

脱敏处理措施，确保处理后的信息无法直接或间接识别特定自然人且不能复原。

（4）合成数据。使用合成数据开展人工智能医学影像研究进行训练或数据增强时，应充分认识其可能存在的伪相关、偏差放大等局限。应明确标注数据的合成属性，并对其生成过程、适用条件及潜在风险进行披露。不得将合成数据混同于真实患者数据而不加说明，不得利用合成数据刻意夸大模型性能或规避伦理审查。

5.2 算法方面

在涉及人工智能医学影像的研究中，应重点关注算法选择、算法训练、模型评估和测试、模型优化与持续迭代中的伦理问题，特别是在使用合成影像数据时可能存在的幻觉、内容失真等问题。

（1）算法安全与可解释性。算法性能（灵敏性、特异性、精确率等）应达到适宜标准，在安全性与透明性上达到医学研究要求；针对深度学习中的算法黑箱，应进行可解释性阐述，确保医生与监管部门能够理解与信任模型的诊断输出。

（2）可复现性与跨环境稳定性。所使用的算法应经过可重复性测试，说明可复现性的条件和特定背景以及必要方法。为确保临床推广的稳健性，应开展跨设备、跨医疗机构、跨人群的复现性验证，评估模型在不同临床环境下的性能稳定性与一致性。

（3）偏见修正。应在数据采集、标注、训练和评估各阶段主动识别潜在偏见，建立偏见量化指标与监测机制，避免数据和

算法中产生的偏见，必要时更新算法、修正偏见。保护算法应用对象，避免偏见转化为对应用对象的歧视。

（4）转化研究的可行性。对算法可信性进行系统验证研究。鼓励向可解释性增强、泛化能力增强、小数据可用性、分布式与联邦学习、多模态融合（影像+病历+基因等）等方向发展。推动算法从实验室研究走向真实世界应用，全面开展内部验证、外部验证、前瞻性临床试验以及多中心真实世界研究，确保模型不仅在技术指标上优越，还能在实际临床场景中为患者带来可衡量的健康获益。

5.3 伦理审查方面

在涉及人工智能医学影像研究的伦理审查中，应重点关注：

（1）合规性、科学价值和社会价值。应重点审查人工智能医学影像研究是否符合国内相关法律法规、国际伦理准则和行业共识，是否以解决临床问题、增进患者福祉、提升公众健康为目的，研究设计与样本估算是否科学合理。

（2）透明可解释、责任可追溯。应审查研究记录（包括可靠性测试与验证过程记录）保存的完整性和清晰性。关注模型在不同设备条件下的泛化能力是否得到充分讨论和验证，以及是否建立完善的追踪机制等。

（3）跟踪审查。鉴于人工智能医学影像研究过程中的风险与不确定性，跟踪审查应重点关注模型上线后的性能漂移及其预警情况，以及迭代后新出现的问题，及时评估与防范应用风险。

（4）数据安全性问题。重点审查数据来源的合法性，确保已获得数据提供方的正式授权；明确划分数据接触与共享的范围及权限；确保数据共享符合可发现、可访问、可互操作和可重用（FAIR）原则并严格遵守相关法规；同时，评估并制定完备的数据安全风险应对措施与应急预案。

**6. 科普宣传**

从事人工智能医学影像研究的科技人员应积极参加伦理培训，提高自身伦理素养，客观准确评价研究成果，避免片面夸大、虚假宣传、误导公众。应积极开展面向研究参与者和社会公众的科学技术与伦理的宣传教育，帮助公众正确认识人工智能医学影像研究的目的和意义，维护医疗信任，推动人工智能技术的健康发展。

本指引由国家科技伦理委员会医学伦理分委员会研究制定，定期评估，适时修订。

国家科技伦理委员会医学伦理分委员会

2026 年 8 月

# 主要参考文件


[1] 中华人民共和国个人信息保护法（2021）

[2] 关于加强科技伦理治理的意见（2022）

[3] 涉及人的生物医学研究伦理审查办法（2016）

[4] 科技伦理审查办法（试行）（2023）

[5] 涉及人的生命科学和医学研究伦理审查办法（2023）

[6] GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型（2019）

[7] GB/T 39576—2020 具有融合功能的移动终端安全能力测试方法（2020）

[8] GB/T 22240—2020 信息安全技术 网络安全等级保护定级指南（2020）.

[9] GB/T 41867—2022 信息技术人工智能术语（2022）

[10] GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范（2025）

[11] 卫生健康行业人工智能应用场景参考指引（2024）

[12] 新一代人工智能治理原则——发展负责任的人工智能（2019）

[13] 新一代人工智能伦理规范（2021）

[14] 核医学名词（定义版）（2018）

[15] 医学影像技术学名词（定义版）（2020）

[16] 自然辩证法名词（第二版、定义版）（2024）

[17] 世界卫生组织卫生健康领域人工智能伦理与治理指南（2021）

[18] 新加坡个人数据保护委员会（PDPC）. Privacy Enhancing Technology (PET): Proposed Guide on Synthetic Data Generation（2024）